JMIR Medical Informatics

Impact Factor (2023): 3.1
Volume 9 (2021), Issue 4 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Review

Original Paper

Reliable Deep Learning–Based Detection of Misplaced Chest Electrodes During Electrocardiogram Recording: Algorithm Development and Validation

Khaled Rjoob[1*], BSc, MSc; Raymond Bond[1*], PhD; Dewar Finlay[1], PhD; Victoria McGilligan[2], PhD; Stephen J Leslie[3], PhD; Ali Rababah[1], MSc; Aleeha Iftikhar[1], PhD; Daniel Guldenring[4], PhD; Charles Knoery[3], MBChB; Anne McShane[5], MSc; Aaron Peace[6], PhD

[1]Faculty of Computing, Engineering & Built Environment, Ulster University, Jordanstown, United Kingdom

[2]Faculty of Life & Health Sciences, Centre for Personalised Medicine, Ulster University, Londonderry, United Kingdom

[3]Department of Diabetes & Cardiovascular Science, University of the Highlands and Islands, Inverness, United Kingdom

[4]HS Kempten, Kempten, Germany, Hochschule Kempten, Kempten, Germany

[5]Emergency Department, Letterkenny University Hospital, Donegal, Ireland

[6]Western Health and Social Care Trust, Londonderry, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Khaled Rjoob, BSc, MSc
Faculty of Computing, Engineering & Built Environment
Ulster University
Shore Road
Jordanstown, BT37 0QB
United Kingdom
Phone: 44 07904392923
Email: rjoob-k@ulster.ac.uk

Abstract

Background: A 12-lead electrocardiogram (ECG) is the most commonly used method to diagnose patients with cardiovascular diseases. However, there are a number of possible misinterpretations of the ECG that can be caused by several different factors, such as the misplacement of chest electrodes.

Objective: The aim of this study is to build advanced algorithms to detect precordial (chest) electrode misplacement.

Methods: In this study, we used traditional machine learning (ML) and deep learning (DL) to autodetect the misplacement of electrodes V1 and V2 using features from the resultant ECG. The algorithms were trained using data extracted from high-resolution body surface potential maps of patients who were diagnosed with myocardial infarction, diagnosed with left ventricular hypertrophy, or a normal ECG.

Results: DL achieved the highest accuracy in this study for detecting V1 and V2 electrode misplacement, with an accuracy of 93.0% (95% CI 91.46-94.53) for misplacement in the second intercostal space. The performance of DL in the second intercostal space was benchmarked with physicians (n=11 and age 47.3 years, SD 15.5) who were experienced in reading ECGs (mean number of ECGs read in the past year 436.54, SD 397.9). Physicians were poor at recognizing chest electrode misplacement on the ECG and achieved a mean accuracy of 60% (95% CI 56.09-63.90), which was significantly poorer than that of DL (P<.001).

Conclusions: DL provides the best performance for detecting chest electrode misplacement when compared with the ability of experienced physicians. DL and ML could be used to help flag ECGs that have been incorrectly recorded and flag that the data may be flawed, which could reduce the number of erroneous diagnoses.

(JMIR Med Inform 2021;9(4):e25347) doi:10.2196/25347

KEYWORDS

deep learning; ECG interpretation; electrode misplacement; feature engineering; machine learning; ECG; engineering; cardiovascular disease; myocardial infarction; myocardial; physicians

Introduction

Background

Clinicians routinely face the challenge of making sense of a large amount of high-dimensional and heterogeneous data to inform their clinical decision making. Poor clinical decisions can fail to provide the correct diagnosis and treatment, which can have a large impact on patient safety and health care costs [1,2]. Artificial intelligence technologies such as deep learning (DL) and machine learning (ML) could play an important role in developing smarter clinical decision-making algorithms that can assist clinicians in making accurate diagnoses. To operationalize artificial intelligence in health care, interactions between data scientists and clinicians are essential to maximize the use of clinical data in the development of automated and predictive systems [2,3].

Cardiovascular diseases are heterogeneous and complex in nature, as they can be caused by a plethora of environmental, genetic, or behavioral factors. To diagnose a cardiac disease, the provision of incorrect data such as electrocardiogram (ECG) data can have a high impact on clinical decision making. A known error is an incorrectly recorded ECG caused by placing precordial electrodes (chest electrodes: V1, V2, V3, V4, V5, and V6) in incorrect positions, resulting in erroneous ECG signals that are interpreted by physicians to inform patient diagnostic signs and treatment plans. This is complicated by the fact that many physicians and cardiologists are not normally present when the ECG is being recorded by a nurse or ECG technician [4-7]. Therefore, ECG interpreters are unaware of the electrode positions that were used to record the ECG that they are reading. Electrode misplacement errors can affect the clinical interpretation of ECGs [8].

Research has shown that signals recorded by electrode V2 are very sensitive to misplacement, followed by electrodes V3, V1, and V4, whereas electrodes V5 and V6 have little visible changes in ECG morphology [9]. The most common error in electrode misplacement is placing electrodes V1 and V2 too high in the third or second intercostal space (ICS). The correct position of electrode V1 is in the fourth ICS at the right sternal edge and that of V2 is in the fourth ICS at the left sternal edge. Correctly placing the electrodes V1 and V2 is crucial, given that their misplacement is also known to cause subsequent misplacement of the remaining chest electrodes (V3 to V6) [9].

Electrode misplacement in ECG acquisition can occur between 40% and 60% of the time [10,11]. Approximately 50% of V1 and V2 electrodes are placed wide and high of their correct anatomical position [10,11].

According to one study, incorrect electrode placement could lead to incorrect diagnoses in 17% to 24% of patients [12]. ECG signals recorded from vertically misplaced V1 and V2 electrodes could also result in a false diagnosis of Brugada syndrome [13] and a failure to detect myocardial infarction (MI) and left ventricular hypertrophy (LVH) [10]. Misplacement can not only conceal but also *mimic* other cardiac diseases, such as MI [14-16]. Less than 20% of cardiologists and 50% of nurses can correctly place V1 and V2 in their correct positions [17]. Several devices have been devised and used to correctly place precordial electrodes. One of the technologies involves using a sliding ruler to facilitate the positioning of electrodes in the correct position [18]. Unfortunately, these technologies have not been widely adopted, likely because of an increase in cost.

To date, research in this area has focused on algorithm development to detect limb electrode interchanges [17-20] rather than precordial electrode misplacement, because the latter is more challenging. Schijvenaars et al [21] used body surface potential maps (BSPMs) to derive transformation matrices to mimic electrode misplacement errors; therefore, BSPMs are suitable for studying electrode misplacement errors.

Objectives

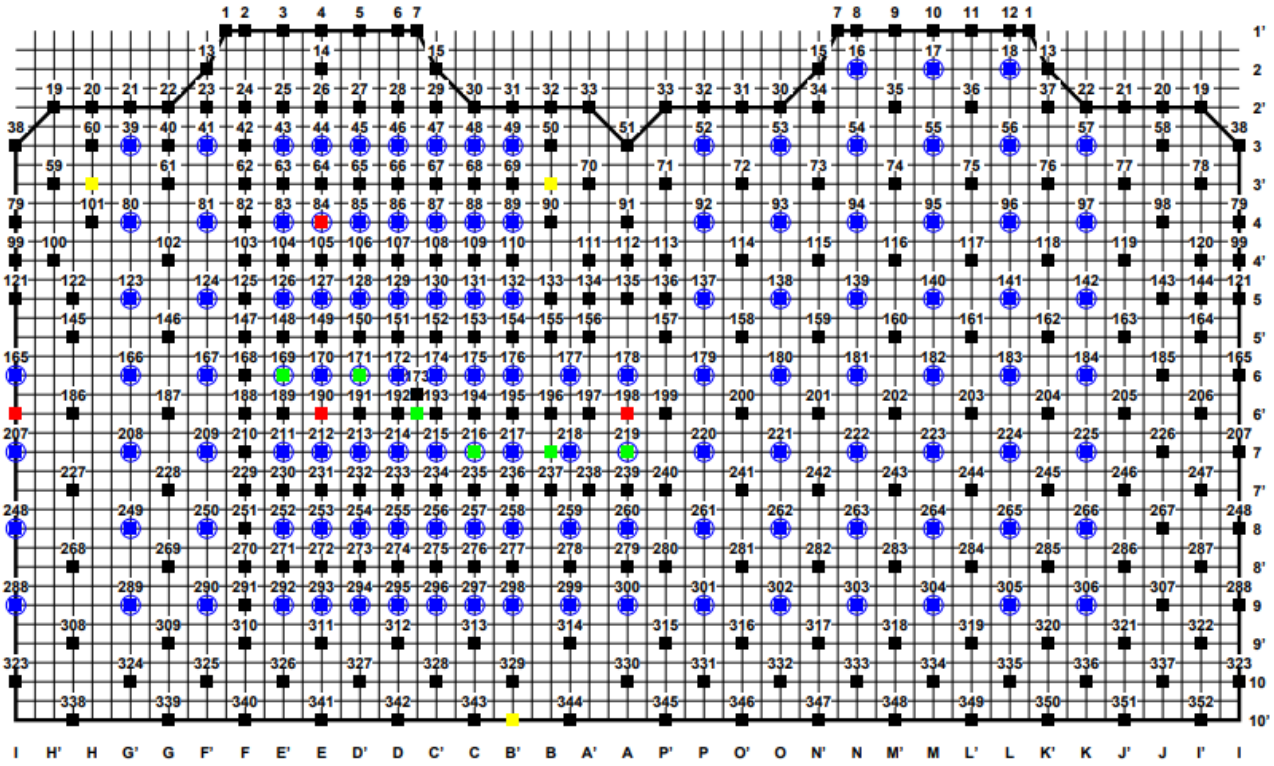
The aim of this study is to determine the performance of ML and DL algorithms for detecting V1 and V2 electrode misplacement when recording ECGs (as V1 and V2 electrode misplacement can also result in misplacement of the other chest electrodes [V3-V6]) and to benchmark this performance against a group of physicians.

Methods

ECG Data Set Description

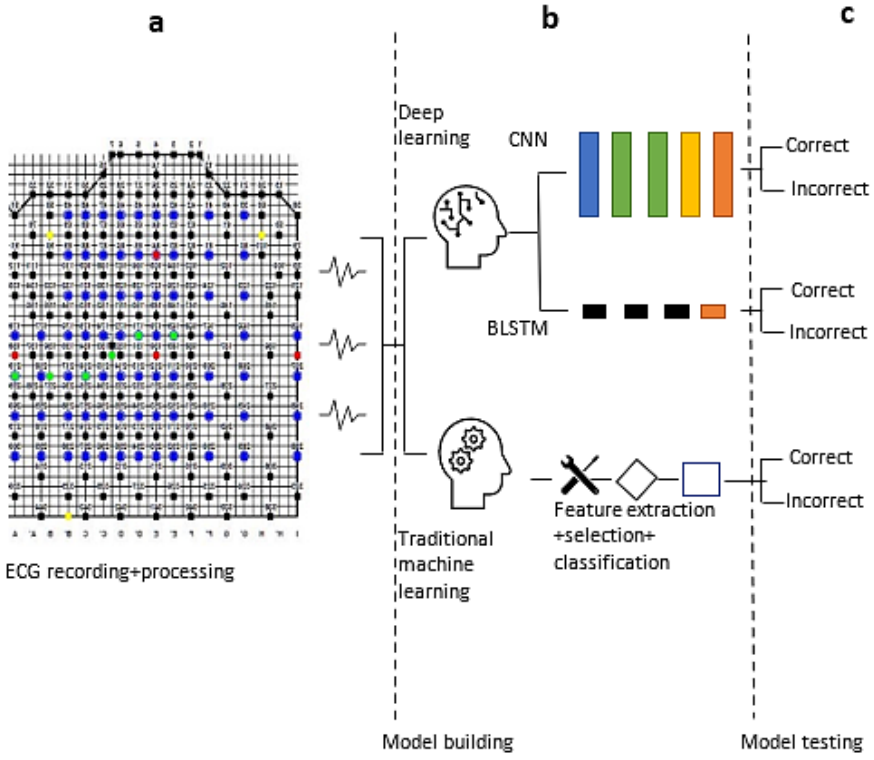
ECGs (V1 and V2 electrodes) were extracted from a high-resolution BSPM (Figure 1). Each BSPM comprises 117 nodes (ECG electrodes) and is known as the Kornreich data set [22-24]. This data set has been used in a large number of publications from groups around the world; however, no researcher has used it to train an algorithm to detect the misplacement of electrodes V1 and V2.

Figure 1. Body surface potential map. Symbols after letters represent column numbers, while symbols after numerals represent row numbers.



The ECG data set consisted of three different subject types, including normal ECGs and ECGs showing MI and LVH. In this study, we have ECGs for 453 patients (normal: n=151, LVH: n=151, and MI: n=151). Each ECG was acquired at a sampling frequency of 300 Hz. For each BSPM, we extracted a correct ECG and an incorrect ECG (where electrodes V1 and V2 were misplaced). This provided a natural class balance where 50% of the cases are correct and 50% are incorrect. This is important given that algorithms improve their performance when being equally exposed to the same number of cases in each class so as to avoid bias and maximize *learning*. For preprocessing, the 117 nodes or electrodes in each BSPM are multiplied using a transformation matrix to obtain 352 nodes that provide greater resolution (using the Dalhousie torso [22], which is a standard approach). According to the recorded data set [25], nodes 169 and 171, denoted in green (Figure 1), represent electrodes V1 and V2, respectively, in their correct positions (fourth ICS). We used nodes 126 and 128, denoted in blue (Figure 1), to represent the misplaced electrodes V1 and V2 in the third ICS and nodes 83 and 85, in blue color, to represent V1 and V2 as misplaced in the second ICS. For each patient, we have recorded the ECG signals simultaneously for electrodes V1 and V2 and one cardiac cycle comprising PQRS deflections. Figure 2 shows an overview of the methodology used in this study.

Figure 2. The data pipeline of this study using 3 phases (data engineering, analytics, and delivery). (a) The data engineering phase that includes data collection (extracting electrocardiograms from body surface potential maps) and data preparation (removing noise from extracted data). (b) The analytics phase that includes traditional machine learning (linear support vector machine, quadratic support vector machine, fine decision tree, coarse decision tree, logistic regression, and bagged tree) and deep learning (convolutional neural network and bidirectional long short-term memory network). (c) The delivery phase that is used to show traditional machine learning and deep learning model inferences. BLSTM: bidirectional long short-term memory; CNN: convolutional neural network; ECG: electrocardiogram.



Detecting V1 and V2 Electrode Misplacement in the Second and Third ICSs Using Feature Engineering

Given that we have one ECG cycle for each patient, the signal was normalized using Equation 1 to reduce signal distortion and baseline drift.

$$y\,[n]=s[n]/\max(|s[n]|) \quad (1)$$

where s[n] is the input signal and y[n] is the output signal.

For feature extraction, a total of 16 ECG features were extracted using 3 different methods: (1) time-domain features (we considered 6 time-domain features, including P-wave amplitude, PR interval, QRS onset value, R-wave peak amplitude, offset of the QRS, and S-wave amplitude), (2) statistical domain features (including the mean, SD, skewness, and kurtosis of the ECG signal; Pearson correlation coefficient; and the root-mean-square error between V1 and V2 electrodes, given that these electrodes are commonly misplaced together), and (3) time-frequency features. The latter involved a discrete wavelet transform using 4 levels and a symlets mother wavelet function, 4 detailed coefficients (D1, D2, D3, and D4), and 4 approximation coefficients (A1, A2, A3, and A4). We also considered the maximum, minimum, and mean values of D4 as features.

For feature selection, a hybrid approach feature selection algorithm was used, which combined the filter and wrapper methods. A total of 16 features were ranked using different filter methods, including mutual information feature selection, maximum relevance minimum redundancy, joint mutual information (JMI), entropy, and relief. Second, a backward elimination algorithm was performed on ranked features to find an optimal set of features as inputs to the ML classifier. The backward elimination algorithm started with all 16 features and removed feature by feature until the best result was achieved.

For classification, we used 6 ML classifiers. This involved the use of three different types of decision trees (DTs): (1) fine DT, which is used to make many leaves that can enable the tree to make fine distinctions between classes; (2) a coarse DT (CDT) that is used to make a small number of leaves that can enable the tree to make coarse distinctions between classes; and (3) a bagged tree that uses bootstrapping with replacement to produce multiple training data sets and takes the majority outcome from multiple trees. Data will be presented using Equation 2.

$$(X;Y)=(x_1,x_2,\ldots\ldots,x_n;\ Y) \quad (2)$$

where X represents features and Y represents classes.

Gini impurity (GI) was the splitting criterion used to split the tree into branches. In this study, there are two classes: (1) label 0 represents the incorrect electrode placement class and (2) label 1 represents the correct electrode placement class. Equation 3 is used to compute the GI for each class.

where n is the number of classes and pi is the fraction of subjects labeled with class i in the data set.

In addition to DT techniques, we used variants of the support vector machine (SVM) and logistic regression (LR). This includes a linear SVM that incorporates two parallel hyperplanes that are selected to separate the data set into two classes where the distance (margin) between hyperplanes should be as large as possible. Equations 4 and 5 describe the two hyperplanes.

$$w.x-b=1 \quad (4)$$

$$w.x-b=-1 \quad (5)$$

where w represents the weight corresponding to each feature, x features the data set, and b represents the biased term. Cases above this hyperplane or on the hyperplane should be in class 1, and cases below this hyperplane should be in class 0.

A quadratic SVM was used, where the quadratic kernel function was used to split the data set into two classes. The difference between linear SVM and quadratic SVM was the kernel function used to split the cases. Finally, LR or logit was used because this was a common statistical technique for binary classification. This technique used log odds (L) as computed using Equation 6, which represents a linear combination of features and model parameters.

$$L= \alpha_0 + \alpha_1.x_1+...+ \alpha_n.x_n \quad (6)$$

where α_0 coefficients are model parameters and xn are features.

Odds (o) computation was the exponent that was used to compute odds using Equation 7, and the corresponding probability was computed using Equation 8.

Detecting V1 and V2 Electrode Misplacement in the Second and Third ICSs Without Feature Engineering

DL does not require feature engineering (ie, feature extraction and selection). Therefore, raw ECG signals are fed into a deep neural network without specifying features. DL can entail different types of networks and architectures. This study uses two different DL networks: (1) convolutional neural networks (CNNs) and (2) bidirectional long short-term memory (BLSTM) networks. A CNN has been built using 15 layers that comprise 1 input layer (used to feed in the ECG signals), 3 hidden convolutional layers (which uses a filter with a variable length to transform the input signal into a convolution layer), 3 batch normalization layers (used to normalize the output of a previous layer by subtracting the mean of batch and dividing this by the SD of the batch), 3 rectified linear unit layers (an activation function that is used to remove negative values), 2 max-pool layers (which combine the sequence output of the previous layer into one single value to reduce the number of parameters and computation in the network), 1 fully connected layer (which connects every neuron in one layer to every neuron in the next layer), 1 soft-max layer (which uses LR to generate probability for each class), and 1 final classification output layer. The BLSTM network comprises 1 sequence input layer, 2 BLSTM hidden layers (which are used to learn the network through each complete time series at each time step), 1 fully connected layer, 1 soft-max layer, and 1 classification output layer.

Physician Detection of V1 and V2 Electrode Misplacements Using Visual Inspection of the ECG

A web-based survey including 30 random ECGs of V1 and V2 (ECGs of correct placement of V1 and V2 [n=15] and ECGs of incorrect placement of V1 and V2 [n=15]) was emailed to 20 participants at the International Society for Computerized Electrocardiology 2019 Conference and Computing in Cardiology 2019 Conference. Of the 20 participants, 11 responded to the survey. They were asked to classify V1 and V2 and whether they were placed correctly. In addition, they were asked about their age, employment status, and experience of reading an ECG (the number of ECGs they read in the past year). A total of 11 physicians responded to the web-based survey. Ethical approval was granted by the Faculty of Computing, Engineering, and Built Environment in Ulster University, Northern Ireland, United Kingdom.

Results

Feature Engineering

As mentioned earlier, 16 features were extracted using three different domains: (1) time domain, (2) statistical domain, and (3) time-frequency domain. Table 1 lists each feature ID along with the feature description. In feature selection (filter process), each feature selection algorithm sorts features from the highest priority feature to the lowest priority feature.

After feature selection, the 6 classifiers were applied, and the best classifier accuracy for detecting misplacement in the second ICS was a bagged DT, followed by CDT, fine DT, LR, quadratic SVM, and linear SVM.

Table 1. Feature IDs and descriptions.

Feature ID	Domain	Feature description
1	Time	P-wave amplitude
2	Time	PR interval
3	Time	QRS beginning value
4	Time	R amplitude
5	Time	End of QRS value
6	Time	S-wave amplitude
7	Statistical	Mean of ECG[a] signal
8	Statistical	Variance of ECG
9	Statistical	SD of ECG signal
10	Statistical	Skewness of ECG
11	Statistical	Kurtosis of ECG signal
12	Time-frequency domain	Maximum value of D4[b]
13	Time-frequency domain	Minimum value of D4
14	Time-frequency domain	Mean value of D4
15	Statistical	Correlation coefficient between V1 and V2 ECGs
16	Statistical	Root-mean-square error between V1 and V2 ECGs

[a]ECG: electrocardiogram.

[b]D4: decomposition coefficient 4.

For detecting electrode misplacement in the third ICS, the best classifier accuracy was also a bagged DT, followed by CDT, LR, quadratic SVM, linear SVM, and fine DT. Table 2 shows the accuracy of each classifier corresponding to each feature selection algorithm. The numeric appended to the label of each feature selection algorithm shows the optimal number of features that was used to achieve the best accuracy. On the basis of classifier accuracy, the best feature selection algorithm performance was JMI for detecting misplacement in the second ICS and RELIEF and JMI for the third ICS.

Table 2. Accuracy of the feature engineering classifiers using machine learning.

Classifier	Percent accuracy in the second ICS[a]					Percent accuracy in the third ICS				
	Entropy15	JMI[b]15	MIFS[c]14	MRMR[d]13	RELIEF16	ENTROPY14	JMI16	MIFS14	MRMR15	RELIEF16
Fine tree	85	85	85	82	84	60	59	60	58	59
Coarse tree	87	87	87	85	87	69	69	69	69	69
Logistic	82	82	83	81	82	64	63	65	63	63
SVM[e]	78	78	75	76	78	59	60	61	61	60
Q-SVM[f]	79	79	78	79	79	58	60	62	60	60
Bagged	88	92	90	92	90	69	70	66	69	70

[a]ICS: intercostal space.

[b]JMI: joint mutual information.

[c]MIFS: mutual information feature selection.

[d]MRMR: maximum relevance minimum redundancy.

[e]SVM: support vector machine.

[f]Q-SVM: quadratic support vector machine.

Superscripts such as MIFS13 represent the best number of features to provide a good accuracy in feature selection algorithm.

DL (Without Feature Engineering)

As mentioned previously, two different DL networks were developed using different architectures. BLSTM achieved the best accuracy compared with the CNN (Table 3) and also outperformed the best accuracy achieved by the

feature-engineered ML classifiers for detecting electrode misplacement in both the second and third ICSs. Figure 3 shows the accuracy, sensitivity, specificity, and receiver operating characteristic curve for BLSTM, CNN, and bagged tree in the second and third ICSs.

Table 3. Classification accuracy using two deep learning networks.

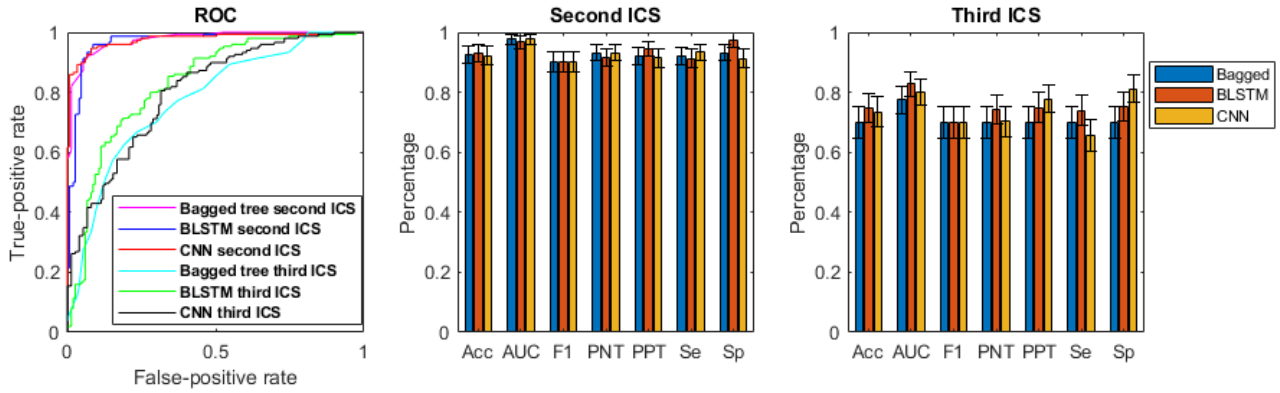
Classifier	Percent accuracy in the second ICS[a]	Percent accuracy in the third ICS
BLSTM[b]	93.0	74.7
CNN[c]	92.3	73.5

[a]ICS: intercostal space.

[b]BLSTM: bidirectional long short-term memory.

[c]CNN: convolutional neural network.

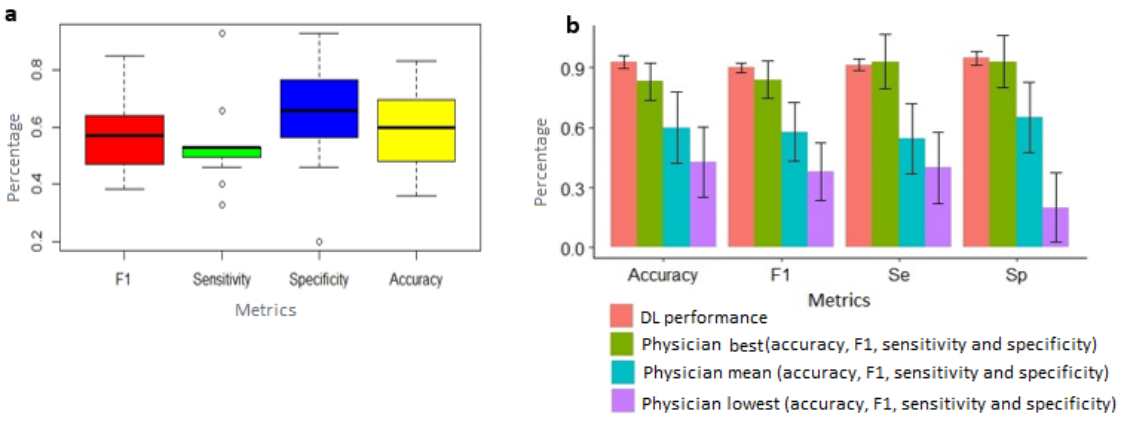
Figure 3. Receiver operating characteristic curves and other metrics results for deep learning and machine learning for detecting electrode misplacement in the second and third intercostal spaces. BLSTM: bidirectional long short-term memory; CNN: convolutional neural network; PNT: predictivity of negative test; PPT: predictivity of positive test.



Physicians Performance in the Second ICS

Performance of 11 physicians (age 47.3, SD 15.5) who were experienced in reading ECGs (mean number of ECGs interpreted in the past year 436.54, SD 397.9) were evaluated using F1 (mean 0.57, SD 0.14), sensitivity (mean 54.5%, SD 15), specificity (mean 65.4%, SD 21), and accuracy (mean 60%, SD 15) when detecting misplacement electrodes V1 and V2 in the second ICS (Figure 4). The accuracy achieved by DL was greater by a factor of 1.5, when compared with the average accuracy of physicians (P<.001).

Figure 4. Physicians' performance for classifying electrocardiograms as correctly recording or as recording with V1 and V2 misplacement in the second intercostal space: (a) physicians' performance and (b) comparison of deep learning performance with physicians' performance regarding different metrics. Error bars were derived using 95% CI (constant=1.96); the pale red bars represent the deep learning performance, whereas the other colors (green, light blue, and purple) represent the physicians' performance (best, mean, and lowest performance, respectively).



Discussion

Principal Findings

On the basis of the medical literature review, ECG electrode misplacement is one of the most critical issues affecting ECG interpretation [8], especially given that it can cause misdiagnoses and inappropriate treatment and potentially a lack of appropriate treatment for the patient. The most common error in chest electrode placement is misplacing electrodes V1 and V2 too high from their correct position that can change ECG morphology and as a result cause a misdiagnosis. In this paper, we present new methods for detecting chest electrode misplacement using 2 approaches: (1) feature-engineered traditional ML algorithms and (2) DL (without any feature engineering) to detect V1 and V2 misplacement. This study describes the first experiment that uses DL to autodetect chest electrode misplacement, whereas previous work mainly focused on limb electrode interchanges. The BLSTM DL network achieved the highest performance in detecting V1 and V2 misplacement in the second ICS with an accuracy of 93.0% and in the third ICS with an accuracy of 74.7%. The ML algorithm (bagged tree) achieved a similar performance with an accuracy of 92.7% (for the second ICS detection) and 70.0% (for the third ICS detection). The performance of the bagged tree and the DL algorithms (BLSTM and CNN) are quite similar, whereas the performance of the other ML algorithms (F tree, C tree, LOG, SVM, and quadratic SVM) is statistically significantly different (P=.01) when compared with the performance of BLSTM, CNN, and bagged tree. A total of 11 medical doctors who were experienced in reading ECGs were recruited to detect electrode misplacement in the second ICS using the same data set to benchmark the ML and DL models. Furthermore, the physicians were biased as they were instructed to identify ECGs that appeared to be recorded incorrectly with respect to the V1 and V2 electrodes. On the basis of their performance, there was a significant difference (P<.001) when compared with the performance achieved by the ML and DL algorithms. Therefore, DL and ML can be used to help flag ECGs that have been incorrectly recorded and flag that the data may be flawed.

More generally, this study is particularly unique as many studies have focused on demonstrating the ability of DL to diagnose patients by automatically interpreting x-rays or ECGs, whereas this study focuses on using DL to detect medical errors. The use of DL to diagnose patients seems to be heavily criticized, given that DL lacks transparency and its decision logic cannot be easily explained to an end user. Therefore, DL for diagnostics elicits many trust issues and may not be widely adopted for this reason. However, physicians may accept black-box systems if they are being used for other subtasks, such as detecting errors, as opposed to providing a patient diagnosis.

Limitations

This study has a number of limitations. The data set is limited and contains only three types of patients (those with MI, LVH, or normal sinus rhythm). Therefore, in further research, new types of patient cases need to be included to increase the data set size and to augment DL performance. Furthermore, the number of participants that manually detected correct or incorrect ECGs was small (n=11), with the limitation being that this cohort may not be a representative sample to benchmark with the ML algorithm. However, the results can be used as a direction for future investigations. The algorithms used were binary in nature and were not tested on many different types of misplacements and variations of ECG recordings. Therefore, a small random variation should be included for all chest electrodes (V1-V6). The performance of the presented algorithms in the real-world setting might not be as accurate as in the study because the algorithm would need to be prospectively tested with patient cases and with different data sets in diverse settings. Moreover, because the misplacement of V1 and V2 can also result in the misplacement of the remaining leads (V3-V6), there is also a need to further understand the impact of the misplacement of V3-V6 electrodes. The performance of the physicians in detecting the misplacement of V1 and V2 electrodes is likely to be lower in the real world as we instructed the subjects to *look out for* and detect the misplacement of V1 and V2 electrodes, which is not likely a condition or a high priority that is at the forefront of a physician's mind when reading an ECG in clinical practice. Given that the ML features used to detect V1 and V2 are somewhat generic, this feature set could be reduced or refined by further clinical insight from experts and the literature that detail V1 and V2 signal morphology when misplaced.

Conclusions

Implementing the algorithms invented in this study could improve ECG data quality, which can, in turn, improve decision making in cardiac care. We can conclude that DL provides the best performance for detecting chest electrode misplacement when compared with ML-based models and the ability of experienced physicians. Therefore, the medical device industry should consider DL to detect chest electrode misplacement. The results clearly show that the greater the misplacement (ie, in the second ICS), the greater the model accuracy. Therefore, in our future research, we aim to improve the accuracy of detecting chest electrode misplacement in the third ICS using alternative techniques rSr' prime. However, adopting these algorithms in health care will take time and will be expensive as it may require prospective testing as part of a trial and approval from different regulatory organizations such as the Food and Drug Administration. However, given that this algorithm is used to flag potential errors and does not provide a diagnosis or recommend treatment, the risks are perhaps less severe. There may still be other costs, including staff training and integrating the algorithm into ECG machines. Future work will also involve the generation of saliency maps that can be used to explain how the DL algorithm is making its decision. This will facilitate knowledge discovery and may result in new ECG features that are characteristic of electrode misplacement.

Acknowledgments

This work was supported by the European Union's INTERREG VA program, managed by the Special EU Programmes Body (SEUPB). The views and opinions expressed in this study do not necessarily reflect those of the European Commission or the SEUPB.

Conflicts of Interest

None declared.

References

1. Neill DB. Using Artificial Intelligence to Improve Hospital Inpatient Care. IEEE Intell. Syst 2013 Mar;28(2):92-95 [FREE Full text] [doi: 10.1109/mis.2013.51]
2. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol 2017 May 30;69(21):2657-2664 [FREE Full text] [doi: 10.1016/j.jacc.2017.03.571] [Medline: 28545640]
3. Yoon J, Davtyan C, van der Schaar M. Discovery and Clinical Decision Support for Personalized Healthcare. IEEE J. Biomed. Health Inform 2017 Jul;21(4):1133-1145 [FREE Full text] [doi: 10.1109/jbhi.2016.2574857]
4. Acampora G, Cook DJ, Rashidi P, Vasilakos AV. A Survey on Ambient Intelligence in Health Care. Proc IEEE Inst Electr Electron Eng 2013 Dec 01;101(12):2470-2494 [FREE Full text] [doi: 10.1109/JPROC.2013.2262913] [Medline: 24431472]
5. Movahedi F, Coyle JL, Sejdic E. Deep Belief Networks for Electroencephalography: A Review of Recent Contributions and Future Outlooks. IEEE J Biomed Health Inform 2018 May;22(3):642-652 [FREE Full text] [doi: 10.1109/JBHI.2017.2727218] [Medline: 28715343]
6. Mortazavi BJ, Desai N, Zhang J, Coppi A, Warner F, Krumholz HM, et al. Prediction of Adverse Events in Patients Undergoing Major Cardiovascular Procedures. IEEE J Biomed Health Inform 2017 Nov;21(6):1719-1729. [doi: 10.1109/JBHI.2017.2675340] [Medline: 28287993]
7. Timmis A. Acute coronary syndromes. BMJ 2015 Oct 20;351:h5153. [doi: 10.1136/bmj.h5153] [Medline: 26487159]
8. Kania M, Rix H, Fereniec M, Zavala-Fernandez H, Janusek D, Mroczka T, et al. The effect of precordial lead displacement on ECG morphology. Med Biol Eng Comput 2014 Feb;52(2):109-119 [FREE Full text] [doi: 10.1007/s11517-013-1115-9] [Medline: 24142562]
9. Rajaganeshan RAA, Ludlam CL, Francis DP, Parasramka SV, Sutton R. Accuracy in ECG lead placement among technicians, nurses, general physicians and cardiologists. Int J Clin Pract 2008 Jan;62(1):65-70. [Medline: 17764456]
10. Wenger W, Kligfield P. Variability of precordial electrode placement during routine electrocardiography. J Electrocardiol 1996 Jul;29(3):179-184. [doi: 10.1016/s0022-0736(96)80080-x] [Medline: 8854328]
11. Bupp JE, Dinger M, Lawrence C, Wingate S. Placement of cardiac electrodes: written, simulated, and actual accuracy. Am J Crit Care 1997 Nov;6(6):457-462. [Medline: 9354224]
12. Bond R, Finlay D, Nugent C, Breen C, Guldenring D, Daly M. The effects of electrode misplacement on clinicians' interpretation of the standard 12-lead electrocardiogram. Eur J Intern Med 2012 Oct;23(7):610-615. [doi: 10.1016/j.ejim.2012.03.011] [Medline: 22939805]
13. Lehmann MH, Brugada R. Brugada syndrome: diagnostic pitfalls. J Emerg Med 2009 Jul;37(1):79-81; author reply 81. [doi: 10.1016/j.jemermed.2008.09.038] [Medline: 19321285]
14. Rudiger A, Schöb L, Follath F. Influence of electrode misplacement on the electrocardiographic signs of inferior myocardial ischemia. Am J Emerg Med 2003 Nov;21(7):574-577. [doi: 10.1016/j.ajem.2003.08.007] [Medline: 14655240]
15. Chanarin N, Caplin J, Peacock A. "Pseudo reinfarction": A consequence of electrocardiogram lead transposition following myocardial infarction. Clin Cardiol 1990 Sep;13(9):668-669. [doi: 10.1002/clc.4960130916] [Medline: 2208827]
16. Jowett NI, Turner AM, Cole A, Jones PA. Modified electrode placement must be recorded when performing 12-lead electrocardiograms. Postgrad Med J 2005 Feb;81(952):122-125 [FREE Full text] [doi: 10.1136/pgmj.2004.021204] [Medline: 15701746]
17. Richard DG. Detecting ECG limb lead-wire interchanges involving the right leg lead-wire. 2017 Computing in Cardiology Conference 2017:1-4 [FREE Full text] [doi: 10.22489/cinc.2017.014-061]
18. Herman MV, Ingram DA, Levy JA, Cook JR, Athans RJ. Variability of electrocardiographic precordial lead placement: a method to improve accuracy and reliability. Clin Cardiol 1991 Jun;14(6):469-476 [FREE Full text] [Medline: 1810683]
19. Xia H, Garcia GA, Zhao X. Automatic detection of ECG electrode misplacement: a tale of two algorithms. Physiol Meas 2012 Sep;33(9):1549-1561. [doi: 10.1088/0967-3334/33/9/1549] [Medline: 22903067]
20. Bond RR, Finlay DD, McLaughlin J, Guldenring D, Cairns A, Kennedy A, et al. Human factors analysis of the CardioQuick Patch®: A novel engineering solution to the problem of electrode misplacement during 12-lead electrocardiogram acquisition. J Electrocardiol 2016;49(6):911-918. [doi: 10.1016/j.jelectrocard.2016.08.009] [Medline: 27662775]
21. Schijvenaars R, Kors J, van Herpen G, van Bemmel J. Use of the standard 12-lead ECG to simulate electrode displacements. Journal of Electrocardiology 1996 Jan;29:5-9 [FREE Full text] [doi: 10.1016/s0022-0736(96)80002-1]

22. Kornreich F, Montague TJ, Rautaharju PM. Identification of first acute Q wave and non-Q wave myocardial infarction by multivariate analysis of body surface potential maps. Circulation 1991 Dec;84(6):2442-2453. [doi: 10.1161/01.cir.84.6.2442] [Medline: 1835677]
23. Montague TJ, Smith ER, Cameron DA, Rautaharju PM, Klassen GA, Felmington CS, et al. Isointegral analysis of body surface maps: surface distribution and temporal variability in normal subjects. Circulation 1981 May;63(5):1166-1172. [doi: 10.1161/01.cir.63.5.1166] [Medline: 7471378]
24. Kornreich F, MacLeod RS, Lux RL. Supplemented standard 12-lead electrocardiogram for optimal diagnosis and reconstruction of significant body surface map patterns. J Electrocardiol 2008;41(3):251-256. [doi: 10.1016/j.jelectrocard.2008.02.011] [Medline: 18433616]
25. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]

Abbreviations

BLSTM: bidirectional long short-term memory
BSPM: body surface potential map
CDT: coarse decision tree
CNN: convolutional neural network
DL: deep learning
DT: decision tree
ECG: electrocardiogram
GI: Gini impurity
ICS: intercostal space
JMI: joint mutual information
LR: logistic regression
LVH: left ventricular hypertrophy
MI: myocardial infarction
ML: machine learning
SEUPB: Special EU Programmes Body
SVM: support vector machine

Edited by C Lovis; submitted 29.10.20; peer-reviewed by R Gregg, A Ramjewan; comments to author 28.12.20; revised version received 12.02.21; accepted 27.02.21; published 16.04.21.

Please cite as:
Rjoob K, Bond R, Finlay D, McGilligan V, J Leslie S, Rababah A, Iftikhar A, Guldenring D, Knoery C, McShane A, Peace A
Reliable Deep Learning–Based Detection of Misplaced Chest Electrodes During Electrocardiogram Recording: Algorithm Development and Validation
JMIR Med Inform 2021;9(4):e25347
URL: https://medinform.jmir.org/2021/4/e25347
doi:10.2196/25347
PMID:33861205

Original Paper

A Framework for Criteria-Based Selection and Processing of Fast Healthcare Interoperability Resources (FHIR) Data for Statistical Analysis: Design and Implementation Study

Julian Gruendner[1], MA, MSc; Christian Gulden[1], MSc; Marvin Kampf[2], MSc; Sebastian Mate[2], Dipl-Inf Univ; Hans-Ulrich Prokosch[1,2], PhD; Jakob Zierk[3], Dr med

[1]Chair of Medical Informatics, Department of Medical Informatics, Biometrics and Epidemiology, Friedrich-Alexander University Erlangen-Nürnberg, Erlangen-Tennenlohe, Germany

[2]Medical Center for Information and Communication Technology, University Hospital Erlangen, Erlangen, Germany

[3]Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany

Corresponding Author:
Julian Gruendner, MA, MSc
Chair of Medical Informatics
Department of Medical Informatics, Biometrics and Epidemiology
Friedrich-Alexander University Erlangen-Nürnberg
Wetterkreuz 15
Erlangen-Tennenlohe
Germany
Phone: 49 91318526785
Email: julian.gruendner@fau.de

Abstract

Background: The harmonization and standardization of digital medical information for research purposes is a challenging and ongoing collaborative effort. Current research data repositories typically require extensive efforts in harmonizing and transforming original clinical data. The Fast Healthcare Interoperability Resources (FHIR) format was designed primarily to represent clinical processes; therefore, it closely resembles the clinical data model and is more widely available across modern electronic health records. However, no common standardized data format is directly suitable for statistical analyses, and data need to be preprocessed before statistical analysis.

Objective: This study aimed to elucidate how FHIR data can be queried directly with a preprocessing service and be used for statistical analyses.

Methods: We propose that the binary JavaScript Object Notation format of the PostgreSQL (PSQL) open source database is suitable for not only storing FHIR data, but also extending it with preprocessing and filtering services, which directly transform data stored in FHIR format into prepared data subsets for statistical analysis. We specified an interface for this preprocessor, implemented and deployed it at University Hospital Erlangen-Nürnberg, generated 3 sample data sets, and analyzed the available data.

Results: We imported real-world patient data from 2016 to 2018 into a standard PSQL database, generating a dataset of approximately 35.5 million FHIR resources, including “Patient,” “Encounter,” “Condition” (diagnoses specified using International Classification of Diseases codes), “Procedure,” and “Observation” (laboratory test results). We then integrated the developed preprocessing service with the PSQL database and the locally installed web-based KETOS analysis platform. Advanced statistical analyses were feasible using the developed framework using 3 clinically relevant scenarios (data-driven establishment of hemoglobin reference intervals, assessment of anemia prevalence in patients with cancer, and investigation of the adverse effects of drugs).

Conclusions: This study shows how the standard open source database PSQL can be used to store FHIR data and be integrated with a specifically developed preprocessing and analysis framework. This enables dataset generation with advanced medical criteria and the integration of subsequent statistical analysis. The web-based preprocessing service can be deployed locally at the hospital level, protecting patients’ privacy while being integrated with existing open source data analysis tools currently being developed across Germany.

(JMIR Med Inform 2021;9(4):e25645) doi:10.2196/25645

KEYWORDS

data analysis; data science; data standardization; digital medical information; eHealth; Fast Healthcare Interoperability Resources; data harmonization; medical information; patient privacy; data repositories; HL7 FHIR

Introduction

Background

With an increase in digitalization in the medical sciences, the efforts to harmonize and standardize clinical data have increased. In particular, transformation of data sets into a common format has received increasing attention to render the data queryable and allow for standardized model building. Two research data repositories with appropriate analysis environments, which have been used extensively and received increasing support, are the OHDSI OMOP common data model [1], which has been designed to facilitate observational research, and Informatics for Integrating Biology and the Bedside (i2b2) [2], which focuses on the integration of different types of data into one clinical repository. Both OHDSI OMOP and i2b2 aim to transform clinical data to a standardized format and vocabulary and are appropriate for research and further analysis. However, importing of data requires the complex implementation of extract, transform, load (ETL) processes [3].

Conversely, the Fast Healthcare Interoperability Resource (FHIR) standard was developed to address the limitations of the previously developed HL7 versions 2 and 3 clinical care document standards; therefore, it is focused on modeling the actual clinical environment as closely as possible. Furthermore, its lightweight nature and direct use of common data formats (ie, JSON and XML) facilitate integration with lightweight webservices. FHIR is now available in its first release with normative resource specifications since version 4.0.0 in 2019 [4], suggesting further maturation of this standard. Large companies including Google, Microsoft, and Apple have adopted FHIR for their medical informatics–related products [5-7]. Moreover, many health system providers are now striving to support or are already supporting the FHIR standard [8], thus potentially facilitating the integration of new solutions into clinical routine, as complex conversions into standards, such as OMOP and i2b2, can be avoided when solutions are deployed within hospitals.

The German Medical Informatics Initiative (MI-I) [9] has recently funded 4 consortia across Germany to investigate how heterogenous clinical data can be integrated into clinical data repositories. One of the objectives of the MI-I is to establish data integration centers (DICs) as the base for cross-hospital and cross-consortia communication. These DICs would provide different services including data integration, data harmonization, standardized data repositories, consent management, and ID management [10-13]. The MI-I has adopted FHIR as the preferred format for inter-consortia communication [14]. All 34 hospitals that are currently part of the MI-I will have a FHIR store available in one form or another and have committed to making their hospital data available in the FHIR format.

The current state of the analysis of FHIR formatted data remains unclear. One drawback of FHIR is that formats such as JSON and XML are not necessarily suitable for further analysis if data are stored in these formats and not processed further. The FHIR standard itself contains an extensive specification for API search operations [15], which, in turn, have their limitations [16]. Specifically, it is not directly possible to express queries with interdata dependencies and necessary computations. Furthermore, searching for resources on the basis of inclusion and exclusion criteria is not possible if they are based on another resource that is not referenced directly but rather indirectly via another intermediary resource. To account for these limitations and to support more complex statistical analysis, a query engine is needed, which should be accessible to researchers without the knowledge of SQL or database query generation and optimization.

Over the years, different FHIR databases have been developed to address the limitations of the FHIR search specification. The blaze FHIR store not only implements the FHIR interface but also introduces the possibility of using clinical quality language to further improve the standard FHIR search and filter possibilities [17]. This platform focuses on feasibility queries and data exports. Another alternative to enhance the availability of FHIR data in an easily accessible manner is to use the PostgreSQL (PSQL) database [18] owing to its innate capability to store, index, and query JSON as binary JSON (jsonb). The fhirbase [19] FHIR database uses PSQL and implements a SQL query interface, which allows a user to query FHIR resources using the SQL syntax. Neither of these solutions currently offer a user-friendly method for a researcher to filter and select data for further statistical analysis, which does not require a strong technical background.

Aim

This study aimed to investigate how a data preprocessing service can be built directly on top of FHIR data stored in a standard PSQL database to enable large data filter queries to generate data sets for statistical analysis. To investigate the requirements for filtering and subset generation, we identified different sample medical data science scenarios. Based on the scenarios' requirements, we defined a data preprocessor, which generates data subsets on the basis of the inclusion and exclusion criteria of other FHIR resources. This data preprocessor was developed to satisfy the demand for investigating subsets of particular FHIR observations and combine them with basic patient data. In this study, which was approved by the institutional review board of the University Hospital Erlangen-Nürnberg (reference# 254_19 Bc), we integrated the developed preprocessing service with a real-world FHIR data set from our hospital, which—at the time of writing—contained approximately 35.5 million FHIR resources. Using this data set, we implemented three different sample medical questions to investigate the capabilities of the implemented web-based preprocessing service. Furthermore, we integrated the service into the locally deployed web-based analysis platform KETOS [20], which enables data retrieval and analysis using Jupyter Notebooks [21] within the hospital, thus respecting a patients' privacy and allowing the data custodians to have ownership of the data.

Methods

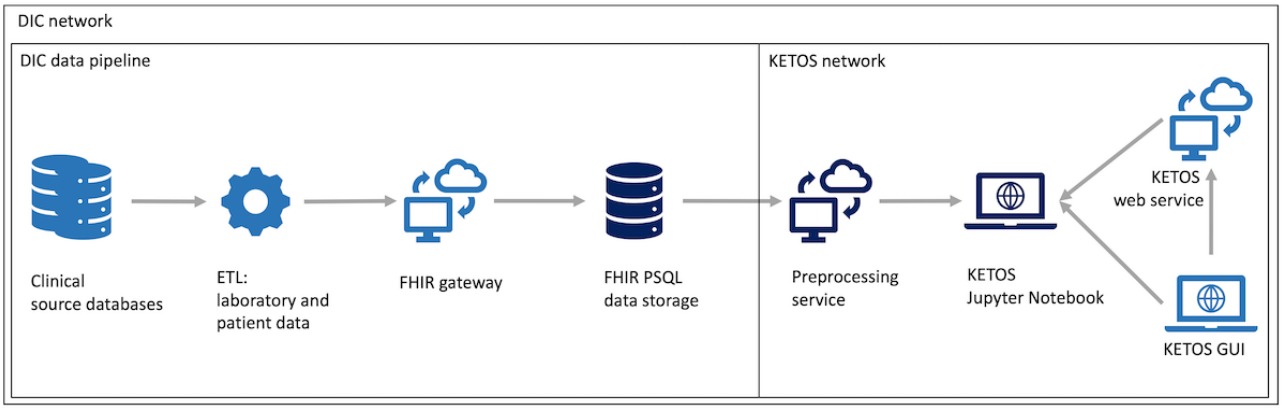
Requirements for a Preprocessing Service

Analysis of large data sets using various statistical methods requires the data to be standardized and harmonized. Additionally, the data set needs to be transformed into a format suitable for further analysis (ie, a "flat structure"). A common approach to this end is to select a subset of data and then convert the selected subset into a simple tabular format. To determine which type of filters are commonly required, we referred to our ongoing multicenter medical research study, wherein we analyzed laboratory findings that are filtered in accordance with the patients' clinical criteria, including patients' diagnoses, clinical procedures, and the results of other laboratory analyses. Specific time criteria (eg, time intervals) are defined for each criterion. This analysis allowed us to identify the following requirements: the ability to select a subset of resources and exclude data from these resources on the basis of their relationship with other resources as inclusion and exclusion criteria. Further, these resources would have to be preprocessed on premises and integrated with the existing DIC infrastructure, so that the analysis could be performed within the hospital on pseudonymized data to adhere to patient privacy and data security regulations in Germany.

Integration With the Existing Infrastructure: the German DIC

We propose that the web-based nature and the reliance on a standard PSQL database with only one table ensures the easy integration of this system into existing infrastructure. Figure 1 shows some components of the DIC infrastructure and some of the analysis tools currently being developed in Germany. The DIC, as currently deployed across 10 German University Hospitals, includes ETL jobs to convert existing data into the FHIR format; moreover, it has a FHIR gateway component, which accepts FHIR resources and loads them into a FHIR PSQL database. This PSQL database, which is the focus of this study, is a standard PSQL database that contains a single table with the following columns: id, fhir_id, type, and data. The data column contains the respective FHIR resource in jsonb format, allowing one to query each element of the JSON stored data directly, while providing complete functionality of a PSQL relational database, like JOINS, timestamp conversion, and LIKE pattern searches. Therefore, a preprocessing service built on this data structure could be run within any hospital as long as the FHIR gateway and the FHIR PSQL database are installed. The entire infrastructure is available in the form of Docker containers and can be easily distributed to other sites. The preprocessing service in this study is web-based and hence integrates well with other web-based platforms for further analysis, such as the KETOS platform for statistical analysis.

Figure 1. Integration with the infrastructure of the data integration center: data storage, preprocessing, and analysis environment. DIC: data integration center, ETL: extract transform and load, FHIR: Fast Healthcare Interoperability Resources, PSQL: PostgreSQL.



The Data Set

The FHIR PSQL database, which we connected our preprocessing service to, contains data on 170,389 patients and 323,779 encounters over 3 years from 2016 to 2018. Among these patients, 88,473 were female, 81,914 were male, and 2 were of an unknown or unspecified gender.

The data sources included the hospital's standardized billing data, which each German hospital is legally required to provide, and laboratory data from a local data warehouse. These data had been harmonized, and laboratory data were mapped to the LOINC vocabulary; diagnoses to International Classification of Diseases, Tenth Revision codes; and procedures to OPS codes. Further, the local DIC pseudonymized the data and harmonized the laboratory units of measurement. The final data set derived from the process included 31,697,035 FHIR Observations, of which 31,686,060 were laboratory findings, 1,740,632 were International Classification of Diseases, Tenth Revision–coded FHIR Conditions, 1,637,573 were FHIR Procedures, 132 were FHIR Medications, and 10,348 were FHIR MedicationSatements. After preprocessing this data set, the final subsets were obtained.

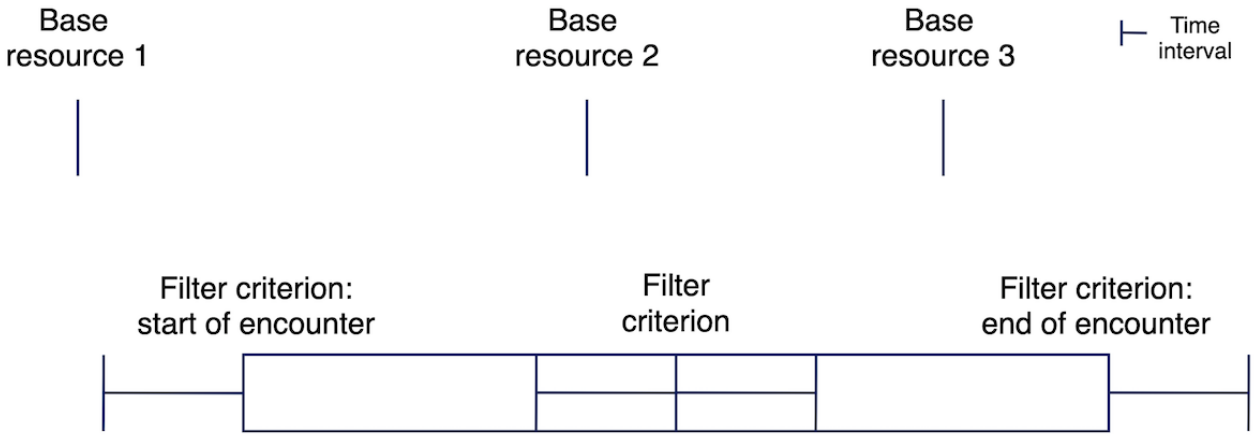
Specification of the Filter Criteria

Through the aforementioned analysis, we established that the preprocessing service should be able to filter all resources from the initial result set (which we referred to as the "base resources"), either on the basis of inclusion or exclusion filter criteria or a combination of both, where a filter criterion is based on another FHIR resource. The filter would then be applied either if a filter criterion ever matched for a patient or if a criterion matched for a patient within a particular time interval of the resources from the base resource to be filtered. Further,

as the time of a laboratory result can often not be directly determined, it was important to determine this period on the basis of the encounter of the filter criterion. If no encounter is available, the laboratory result is filtered on the basis of the criterion along with a time interval. The resulting logic for resource matching based on the time from the base resource is depicted in Figure 2. The figure shows three base resources: 1, 2, and 3. Resource 1 would not be filtered from the result set because it lies outside the specified time interval. Resource 2 would be filtered because it lies within the specified time interval from the filter criterion. Resource 3 would be filtered if the filter criterion has an encounter because it lies within the time interval of the encounter of the filter criterion. However, it would not be filtered if the filter criterion does not have an encounter.

In addition to the possibility of defining the time interval within which a resource must be filtered, filter criteria should be selected on the basis of their respective code (eg, LOINC code 718-7 for hemoglobin). Further, it should be possible to specify a simple value restriction in accordance with standard comparators.

Figure 2. Timeline for filter matching.



Data Availability

The source code of the project is available on GitHub [22].

Results

Overview of the Findings

We implemented and deployed the preprocessing service that we implemented in this study at the University Hospital Erlangen-Nürnberg. The whole pipeline could be easily deployed on an existing server, as the web-based preprocessing service was packaged as a Docker container [23].

To ensure secure functioning of the preprocessing service, we deployed it on the same server and within the same Docker network as the KETOS analysis environment. The preprocessing service was then only made available within the Docker network on the server and was not accessible outside the KETOS platform. Finally, we applied the preprocessing service to patient data from 2016 to 2018 stored in the local DIC FHIR database (see *The Data Set*). We used the preprocessing service to generate 3 sample data sets and analyses to demonstrate its applicability to clinically relevant research questions. We then analyzed the resulting prepared data sets with the KETOS platform and a Jupyter Notebook (interactive cell-based code development in a web browser).

Specification of the Preprocessing Service and Data Input

Based on the data analysis and the specification of filter criteria, we described an interface that receives the input parameters in JSON format (Multimedia Appendix 1) and uses the input to generate a PSQL filter query (Multimedia Appendix 2), which is sent to the FHIR PSQL database where the query is executed. This query yields a subset of resources. The preprocessor then generates a feature set using this subset and combines the subset with basic patient data to generate the final feature data set for further statistical analysis, as specified in the feature_set part of the input parameter JSON. The initial filtered resource set and the final feature set are then stored in the preprocessors' own local database ready to be downloaded for analysis. The preprocessor itself was implemented as a webservice, using the Python Flask-Restful library [24].

Example 1: Data-Driven Establishment of Reference Intervals

In modern medicine, laboratory tests are an essential tool for health assessment and substantially influence diagnostic and treatment decisions. To support decision making among clinicians, laboratory findings are accompanied by reference intervals, which reflect the range of test results in a population of healthy individuals. Conventionally, reference intervals have been established among specifically recruited healthy individuals ("direct approach"); however, this approach is associated with substantial financial and logistical challenges. Therefore, data-driven approaches ("indirect approaches") have been developed, which use data from laboratory information systems and statistical analyses to estimate the proportion of samples from healthy individuals in mixed data sets (ie, hospital data sets containing a large fraction of abnormal test results). While indirect approaches can tolerate a high proportion of abnormal

findings, their accuracy is limited by the proportion of abnormal samples.

Here, we demonstrate how reference intervals for a very common laboratory test (hemoglobin) can be established using the tools developed in this study and the open source kosmic [25] algorithm, despite a very high proportion of abnormal findings in the analyzed data set (ie, in-patient laboratory findings from a tertiary care center). To reduce the number of abnormal findings, we excluded all patients with cancer diagnoses (defined by ICD codes starting with *C*) and those who received transfusions (defined using OPS codes starting with *8-80*) at any time. Additionally, we excluded all findings from patients with clearly abnormal hemoglobin values (ie, <8.0 g/dL) at any time and those having undergone surgery within 90 days (defined by OPS codes starting with *5*-). We then restricted the data set to a sample of interest (men aged 18-65 years), selected one random finding per patient and used the resulting data set (n=13,721) as input for the kosmic algorithm. This yielded a clinically useful reference interval (13.2-17.2 g/dL), which highlights the potential of the developed framework to handle complex medical data science scenarios.

Example 2: Anemia in Patients With Cancer

Assessment of differences in laboratory findings among different patient cohorts enhances physicians' understanding of the pathophysiology of diseases and treatment effects. To assess the feasibility of such analyses using the tools developed in this study, we generated a data set to investigate anemia occurrence (ie, hemoglobin levels below cut-off values defined by the World Health Organization) among adult patients with and those without cancer. We queried the minimum hemoglobin level (defined using LOINC code *718-7*) of patients with and those without cancer (included or excluded using ICD codes starting with *C*) and determined the number of adult patients below anemia-defining thresholds (13 g/dL for men and 12 g/dL for women). In total, this resulted in a data set with 686,472 hemoglobin test results from 9075 men and 9035 women with cancer and 45,766 men and 53,777 women without cancer. We observed a substantially larger proportion of men and women with anemia among patients with cancer (n=6316, 69.6% and n=5674, 62.8%, respectively) than among those without cancer (n=16,247, 35.5% and n=22,586, 42.0%, respectively) (P<.001, Fisher exact test). These findings indicate a high prevalence of anemia, a condition associated with substantial morbidity and mortality, in cancer (ie, the second most common cause of death worldwide) and the suitability of the tools developed in this study for such analyses.

Example 3: Adverse Effects of Drugs

Adverse effects of drugs are a major contributor to patient morbidity and mortality among in-hospital patients and outpatients, and a substantial proportion of drugs' adverse effects influence laboratory findings. Here, we used the framework developed in this study to generate a data set to investigate changes in patients' potassium levels during treatment with an important anti-infective drug (liposomal amphotericin B, a potent and essential antifungal agent that decreases potassium levels in some patients). We selected potassium levels (defined using LOINC code *2823-3*) in patients who received liposomal amphotericin B (defined using OPS codes starting with *6-002.q*) within 7 days (study group: 107 patients and 4568 potassium test results) and potassium levels in patients who received liposomal amphotericin B at any time but not within 7 days (control group: 145 patients and 5581 potassium test results). This example shows that this framework can be used to generate a data set to investigate the adverse effects of drugs. Although potassium levels did not significantly differ between both groups in this data set (P=.12), they were lower in the study group (3.4 mM) than in the control group (3.5 mM), demonstrating the ability of this framework to investigate the adverse effects of drugs.

Discussion

Principal Findings

Direct retrieval of data stored using FHIR resources for further statistical analysis is an important step to bridge the gap between the acquisition of medical data and clinically relevant research. To comply with patient privacy and data security regulations, it is important to establish tools that can be directly deployed within the hospital infrastructure, so that the data remain within the institutions' network and control. The preprocessor we developed satisfies these concerns and relies on open source tools that can be easily distributed across hospitals to improve future research. Further, since this preprocessor relies on FHIR resources, extra ETL jobs converting the FHIR clinical data format—which is currently supported directly by vendors of electronic health records into other data storage formats such as OMOP and i2b2—are unnecessary. The largest challenge for the FHIR standard is the ability to use the data for further analysis. Nonetheless, even research-driven formats such as OMOP and i2b2 often need further processing for detailed statistical analysis. For example, for further data analysis using DataSHIELD, a distributed privacy preserving data analysis platform, further processing of OMOP and i2b2 data is necessary [26]. This indicates that direct processing of FHIR resources can reduce the overall complexity and help avoid extra transformation steps.

This study shows that the use of PSQL to store FHIR data and further build web-based preprocessors on this infrastructure is a viable way to handle large amounts of clinical data without having to rely on cloud-based or proprietary data storage solutions. This not only retains a hospital's ownership of its data but also allows the hospital to avoid vendor lock-in. Development of the preprocessor as a webservice implies that integration into web-based tools can be easily achieved, and the generation of a web-based JavaScript user interface, for example, can be inherently supported. The tool developed in this study does not require the FHIR data to be harmonized across hospitals; however, cross-hospital data analysis is only viable if data are harmonized. Direct integration into the DIC infrastructure developed across Germany and the DIC ensuring data harmonization, including LOINC mapping for laboratory values and LOINC harmonization and unit harmonization through conversion, would facilitate future multicenter studies. Using 3 clinically meaningful scenarios and a real-world data

set, we demonstrated the usefulness of the developed framework .

This study integrated the preprocessing service with the KETOS environment and directly interacted with the preprocessor from within a Jupyter Notebook. We made the preprocessor available only within the KETOS platform, allowing it to be password protected by default. Deployment of this platform within a hospital—after pseudonymizing the data and confining it to the hospital for further analysis—ensures patient privacy. Specifically, this framework facilitates retrospective analyses of large data sets, where consent for the data to leave the hospital confines cannot be reasonably obtained. Duplication of this framework across institutions allows data custodians and researchers within each institute to perform analyses and then collaborate with researchers from other institutions. The prerequisite for this is that only aggregated data leave the confines of each institution for the final analysis. In a potential workflow, researchers can establish the preprocessing specifications and analysis scripts with a Jupyter Notebook at their institution and share them with collaborators. This allows them to not only check and execute the scripts at their institution but also modify the scripts per their data requirements, if necessary. Aggregated results or generated models can then be shared across the collaborating institutions. Throughout the process, the FHIR format and identical preprocessing ensures that the scripts and specifications are applicable across the institutions.

It is important to note that the preprocessor only generates SQL queries and does not have large hardware requirements because search and filtering are carried out by the well-established open source PSQL database. A more detailed performance test of the implementation is beyond the scope of this study because performance largely depends on database optimization and indexing, and the number of resources identified for the base filter criteria. However, even the longest requests to generate our sample datasets took minutes rather than hours, despite only creating basic indices for resource types and IDs.

Lessons Learned

The development of a preprocessor based on FHIR data stored in PSQL jsonb databases for statistical analysis is a viable alternative, facilitating more advanced data processing when compared to the FHIR Search specified as part of the FHIR standard. The FHIR format itself is suitable for querying because JSON queries can be used to specify preprocessing input parameters. The performance of the PSQL database is limited insofar as handling of large data is strongly influenced by how well the PSQL database is administered. For the database we used in this study, we defined some simple indices on the basis of the fhir_id and the resource type to improve the query performance. Here, we first attempted to implement the preprocessor directly on a FHIR server; however, we found that the HAPI FHIR server did not perform well with large bulk loads, which led to the DIC switching from the HAPI FHIR server to the PSQL database. Therefore, large amounts of data were never directly available in a FHIR server. More complex queries, including pattern searches and combining of data for filtering across resources, were not directly supported by the HAPI server. The initial implementation of the preprocessor based on the HAPI server first downloaded the necessary resources to be processed within the preprocessor; however, this was less efficient than direct processing of the data on the database side. The current implementation focuses on feature selection, wherein one particular feature is selected, and inclusion and exclusion criteria are based on the sought-after feature in relation to other data. A cohort selection process could be implemented by selecting the distinct patient IDs in the result set. A future version of this platform should investigate how these concerns could be separated. A feature selection module can then be built on top of a cohort selection module in a 2-step process.

Generalizability and Use in Other Studies

Reliance on the FHIR format and, more specifically, on fields within the FHIR resources, which are usually set, implies that the proposed method is applicable in various scenarios without requiring further ETL jobs. The preprocessor could process any combination of Observations, Procedures, and Conditions identified by their code within the respective vocabulary. The implementation is currently restricted to the filtering of individual base resources, implying that the generation of data sets where multiple resources are associated with one another based on groups is currently not supported. One could envision an extension, which combines the results of multiple queries into one data set in the future, allowing for more complex analysis. The current version will support the extraction and investigation of any single feature in relation to others. In this study, we demonstrate the investigation of, for example, hemoglobin levels. Any other laboratory value, condition, or procedure would be supported by the current platform. In particular, the method proposed here allows one to filter each occurrence of a feature individually, implying that one query can filter individual occurrences of a feature over time. This facilitates queries, such as the search for hemoglobin value observations around which no blood transfusions have occurred.

Limitations

The preprocessor specified and implemented in this study was developed on the basis of one projects' requirement on data handling. Although this study demonstrates its applicability in various scenarios, it does not satisfy more advanced query mechanisms including those developed by , for example, the OMOP OHDSI group. For instance, this framework lacks deeper temporal logic [27], such as temporal filters (eg, the first observation after a certain event). Furthermore, it is important to note that the preprocessor cannot be directly used to define patient cohorts and feasibility queries because it focuses on extracting one feature in relation to others over time. While this restricts the use of the tool, it allows for more specific identification of individual feature occurrences in relation to others. The preprocessor implemented here cannot provide the extent of out-of-the-box analysis which the OMOP and i2b2 tool suites provide; however, it clearly demonstrates the feasibility of building preprocessing tools for FHIR-formatted data. Overall, the data selection and extraction processes specified here have to be used in combination with analysis tools such as DataSHIELD or Jupyter Notebooks, allowing

researchers to apply use-case–specific analysis tools to the extracted data or, in the case of DataSHIELD, use the data sets for cross-hospital analysis.

Further, this preprocessing service is dependent on JSON input, and it lacks a user interface. Finally, building on top of a PSQL database restricts the preprocessor to the PSQL database, which implies that some of the interoperability that the FHIR standard aims at is lost in the process, and the current solution cannot replace an FHIR server. However, as the data has to be transformed for analysis regardless, it still provides a viable alternative for FHIR data storage for further analysis.

Future Directions

This study shows that PSQL jsonb lends itself well to being extended with preprocessing services for data modeling. Further studies are required to investigate how to create a preprocessing tool for the FHIR format, which has similar capabilities to those of the OHDSI OMOP ATLAS or the i2b2 querying tools. In this pursuit, studies should evaluate whether the existing tools already implement all necessary logic for developing and analyzing statistical models. The preprocessor developed here currently lacks a user interface, which is an important requirement for any preprocessor to make it more accessible to a wider audience with different technical backgrounds. We recommend the development of a user interface as an important subsequent step while simultaneously improving this preprocessor. Furthermore, studies should investigate how well different FHIR databases lend themselves to advanced processing of data needed to generate a data set for statistical analysis. For practical reasons (ie, the data being available in our consortium in a simple PSQL database containing one table), we built the preprocessor on top of this PSQL schema. Depending on the outcome of the analysis of the available FHIR stores, a cohort and feature selection mechanism could be developed on the fhirbase project or other solutions, including the clinical quality language capable blaze FHIR store or an extended FHIR search specification and implementation. Criteria-based resource selection is only a small part of a larger analysis framework, similar to OHDSI OMOP and i2b2, which is currently missing for the FHIR standard and should be developed in the future. However, even for larger data sets, direct preprocessing on FHIR resources is a feasible alternative and should be further investigated.

Conclusion

The preprocessor developed in this study demonstrates how standard open source tools including PSQL can be used to store FHIR data in a format that can be used to generate further filtering, cohort, and feature selection mechanisms. We further deployed the tool at the University Hospital Erlangen-Nürnberg and applied the preprocessor to a large pool of data, generated 3 sample data sets, and executed analyses on top of the generated data sets to demonstrate the applicability of this preprocessor in research. These queries included multiple FHIR resources, such as Observation, Condition, Procedure, Patient, and Encounter, demonstrating the capability of our implementation.

Acknowledgments

This study was performed in fulfilment of the requirements for obtaining the degree “Dr rer Biol Hum” at the Friedrich-Alexander Universität Erlangen-Nürnberg (JG). This study was supported by grants from the MIRACUM project (which is funded by the German Ministry for Education and Research; funding# FKZ: 01ZZ1801A) and has been conducted within the MIRACUM consortium.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Example query json specification.
[PDF File (Adobe PDF File), 75 KB - medinform_v9i4e25645_app1.pdf]

Multimedia Appendix 2
Example query generated sql.
[PDF File (Adobe PDF File), 89 KB - medinform_v9i4e25645_app2.pdf]

References

1. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
2. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124-130 [FREE Full text] [doi: 10.1136/jamia.2009.000893] [Medline: 20190053]
3. Maier C, Lang L, Storf H, Vormstein P, Bieber R, Bernarding J, et al. Towards Implementation of OMOP in a German University Hospital Consortium. Appl Clin Inform 2018 Jan;9(1):54-61 [FREE Full text] [doi: 10.1055/s-0037-1617452] [Medline: 29365340]

4. Welcome to FHIR. HL7 International. URL: https://www.hl7.org/fhir/ [accessed 2020-02-04]
5. Cloud Healthcare API Internet. Google Cloud. URL: https://cloud.google.com/healthcare/ [accessed 2020-04-23]
6. Azure API for FHIR Internet. Microsoft Azure. URL: https://azure.microsoft.com/en-us/services/azure-api-for-fhir/ [accessed 2020-04-23]
7. Accessing Health Records. Apple Developer. URL: https://developer.apple.com/documentation/healthkit/samples/accessing_health_records [accessed 2021-02-26]
8. Posnack S, Barker W. Heat Wave: The U.S. is Poised to Catch FHIR in 2019. Health IT Buzz. 2018 Oct 01. URL: https://www.healthit.gov/buzz-blog/interoperability/heat-wave-the-u-s-is-poised-to-catch-fhir-in-2019 [accessed 2020-02-26]
9. Semler S, Wissing F, Heyder R. German Medical Informatics Initiative. Methods Inf Med 2018 Jul;57(S 01):e50-e56 [FREE Full text] [doi: 10.3414/ME18-03-0003] [Medline: 30016818]
10. Winter A, Stäubert S, Ammon D, Aiche S, Beyan O, Bischoff V, et al. Smart Medical Information Technology for Healthcare (SMITH). Methods Inf Med 2018 Jul;57(S 01):e92-e105 [FREE Full text] [doi: 10.3414/ME18-02-0004] [Medline: 30016815]
11. Haarbrandt B, Schreiweis B, Rey S, Sax U, Scheithauer S, Rienhoff O, et al. HiGHmed - An Open Platform Approach to Enhance Care and Research across Institutional Boundaries. Methods Inf Med 2018 Jul;57(S 01):e66-e81 [FREE Full text] [doi: 10.3414/ME18-02-0002] [Medline: 30016813]
12. Prasser F, Kohlbacher O, Mansmann U, Bauer B, Kuhn K. Data Integration for Future Medicine (DIFUTURE). Methods Inf Med 2018 Jul;57(S 01):e57-e65 [FREE Full text] [doi: 10.3414/ME17-02-0022] [Medline: 30016812]
13. Prokosch H, Acker T, Bernarding J, Binder H, Boeker M, Boerries M, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. Methods Inf Med 2018 Jul;57(S 01):e82-e91 [FREE Full text] [doi: 10.3414/ME17-02-0025] [Medline: 30016814]
14. Medizininformatik-Initiative beschließt Verwendung von FHIR. Medical Informatics Initiative Germany. URL: https://www.medizininformatik-initiative.de/en/node/312 [accessed 2020-04-24]
15. Search. HL7 International. URL: https://www.hl7.org/fhir/search.html [accessed 2020-06-15]
16. Gulden C, Mate S, Prokosch H, Kraus S. Investigating the Capabilities of FHIR Search for Clinical Trial Phenotyping. Stud Health Technol Inform 2018;253:3-7. [Medline: 30147028]
17. Samply/Blaze. GitHub. URL: https://github.com/samply/blaze [accessed 2020-05-12]
18. PostgreSQL: The World's Most Advanced Open Source Relational Database Internet. The Postgresql Global Development Group. URL: https://www.postgresql.org/ [accessed 2021-02-26]
19. Fhirbase. GitHub. URL: https://github.com/fhirbase/fhirbase [accessed 2020-05-12]
20. Gruendner J, Schwachhofer T, Sippl P, Wolf N, Erpenbeck M, Gulden C, et al. Correction: KETOS: Clinical decision support and machine learning as a service - A training and deployment platform based on Docker, OMOP-CDM, and FHIR Web Services. PLoS One 2019;14(11):e0225442 [FREE Full text] [doi: 10.1371/journal.pone.0225442] [Medline: 31721815]
21. Jupyter. Project Jupyter. URL: http://jupyter.org [accessed 2021-02-26]
22. criteria-based-selection-fhir. GitHub. 2020 Nov 09. URL: https://github.com/juliangruendner/criteria-based-selection-fhir [accessed 2021-03-02]
23. Get Started with Docker. Docker. URL: https://www.docker.com/ [accessed 2020-02-04]
24. Flask-RESTful. Flask-RESTful. URL: https://flask-restful.readthedocs.io/en/latest/ [accessed 2021-01-29]
25. Zierk J, Arzideh F, Kapsner LA, Prokosch H, Metzler M, Rauh M. Reference Interval Estimation from Mixed Distributions using Truncation Points and the Kolmogorov-Smirnov Distance (kosmic). Sci Rep 2020 Feb 03;10(1):1704 [FREE Full text] [doi: 10.1038/s41598-020-58749-2] [Medline: 32015476]
26. Horki P, Lenz S, Gruendner J, Maier C, Alexander L, Boeker M. omopRds: transfer of data models from OMOP to DataSHIELD/Opal Internet. 2019 Presented at: 64. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS); Dortmund; Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie; 2019; Düsseldorf, Germany URL: https://www.egms.de/static/de/meetings/gmds2019/19gmds028.shtml [doi: 10.3205/19gmds028]
27. Mate S, Bürkle T, Kapsner LA, Toddenroth D, Kampf MO, Sedlmayr M, et al. A method for the graphical modeling of relative temporal constraints. J Biomed Inform 2019 Dec;100:103314 [FREE Full text] [doi: 10.1016/j.jbi.2019.103314] [Medline: 31629921]

Abbreviations

DIC: data integration center
ETL: extract, transform, load
FHIR: Fast Healthcare Interoperability Resources
i2b2: Informatics for Integrating Biology and the Bedside
JSON: JavaScript Object Notation
jsonb: binary JavaScript Object Notation
MI-I: German Medical Informatics Initiative

PSQL: PostgreSQL

Edited by C Lovis; submitted 10.11.20; peer-reviewed by N Sakib, JY Kim; comments to author 16.01.21; revised version received 29.01.21; accepted 31.01.21; published 01.04.21.

Please cite as:
Gruendner J, Gulden C, Kampf M, Mate S, Prokosch HU, Zierk J
A Framework for Criteria-Based Selection and Processing of Fast Healthcare Interoperability Resources (FHIR) Data for Statistical Analysis: Design and Implementation Study
JMIR Med Inform 2021;9(4):e25645
URL: https://medinform.jmir.org/2021/4/e25645
doi:10.2196/25645
PMID:33792554

Original Paper

Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer Using Harmonized Observational Databases: Retrospective Study

Hokyun Jeon[1*], MS; Seng Chan You[2,3*], MD, MS; Seok Yun Kang[4], MD; Seung In Seo[5], MD; Jeremy L Warner[6], MD, MS; Rimma Belenkaya[7], MA, MS; Rae Woong Park[1,2], MD, PhD

[1]Department of Biomedical Sciences, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea

[2]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea

[3]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[4]Department of Hematology-Oncology, Ajou University School of Medicine, Suwon, Gyeonggi-do, Republic of Korea

[5]Department of Internal Medicine, Kangdong Sacred Heart Hospital, Hallym University College of Medicine, Seoul, Republic of Korea

[6]Division of Hematology and Oncology, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States

[7]Department of Health Informatics, Memorial Sloan Kettering Cancer Center, New York, NY, United States

[*]these authors contributed equally

Corresponding Author:
Rae Woong Park, MD, PhD
Department of Biomedical Informatics
Ajou University School of Medicine
Hong Jae Gwan, 5th Fl
164, World cup-ro, Yeongtong-gu
Suwon, Gyeonggi-do, 16499
Republic of Korea
Phone: 82 31 219 4471
Email: rwpark99@gmail.com

Abstract

Background: Accurate and rapid clinical decisions based on real-world evidence are essential for patients with cancer. However, the complexity of chemotherapy regimens for cancer impedes retrospective research that uses observational health databases.

Objective: The aim of this study is to compare the anticancer treatment trajectories and patterns of clinical events according to regimen type using the chemotherapy episodes determined by an algorithm.

Methods: We developed an algorithm to extract the regimen-level abstracted chemotherapy episodes from medication records in a conventional Observational Medical Outcomes Partnership (OMOP) common data model (CDM) database. The algorithm was validated on the Ajou University School Of Medicine (AUSOM) database by manual review of clinical notes. Using the algorithm, we extracted episodes of chemotherapy from patients in the EHR database and the claims database. We also developed an application software for visualizing the chemotherapy treatment patterns based on the treatment episodes in the OMOP-CDM database. Using this software, we generated the trends in the types of regimen used in the institutions, the patterns of the iterative chemotherapy use, and the trajectories of cancer treatment in two EHR-based OMOP-CDM databases. As a pilot study, the time of onset of chemotherapy-induced neutropenia according to regimen was measured using the AUSOM database. The anticancer treatment trajectories for patients with COVID-19 were also visualized based on the nationwide claims database.

Results: We generated 178,360 treatment episodes for patients with colorectal, breast, and lung cancer for 85 different regimens. The algorithm precisely identified the type of chemotherapy regimen in 400 patients (average positive predictive value >98%). The trends in the use of routine clinical chemotherapy regimens from 2008-2018 were identified for 8236 patients. For a total of 12 regimens (those administered to the largest proportion of patients), the number of repeated treatments was concordant with the protocols for standard chemotherapy regimens for certain cases. In addition, the anticancer treatment trajectories for 8315 patients were shown, including 62 patients with COVID-19. A comparative analysis of neutropenia showed that its onset in colorectal cancer regimens tended to cluster between days 9-15, whereas it tended to cluster between days 2-8 for certain regimens for breast cancer or lung cancer.

Conclusions: We propose a method for generating chemotherapy episodes for introduction into the oncology extension module of the OMOP-CDM databases. These proof-of-concept studies demonstrated the usability, scalability, and interoperability of the proposed framework through a distributed research network.

(JMIR Med Inform 2021;9(4):e25035) doi:10.2196/25035

KEYWORDS

antineoplastic combined chemotherapy protocols; electronic health record; cancer; pattern; chemotherapy; database; retrospective; algorithm; scalability; interoperability

Introduction

Background

In cancer research, real-world data, with the exception of cancer registries, have been relatively underused despite recent advances in information technology and the availability of data from electronic health records (EHRs) or administrative claims databases [1]. One of the major challenges to the active use of EHRs or claims databases in cancer research is the limited availability of clinically relevant structured data elements. The essential data elements for cancer research include records of anticancer treatment at the unit of the regimen, rather than the individual drugs used in the regimen. In order for researchers to obtain the data necessary for conducting a comparative study of the treatment regimens of patients with cancer, a labor-intensive preprocess is inevitable [2-4].

Prior Work

Previously, researchers developed algorithms to replace the manual endeavor of capturing the details of chemotherapy from medication histories [5-8]. Even though these studies carved paths toward the use of real-world evidence in cancer research, none of them focused on identifying and addressing organizational barriers. Due to heterogeneity in the structure and semantics of EHRs or claims databases across institutions and countries, none of these studies provided a scalable framework.

The Observational Health Data Sciences and Informatics (OHDSI) collaboration, which is a multistakeholder group organized for global collaboration studies, provides the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to contribute to medical research on harmonized observational databases [9]. The oncology work group in the OHDSI community has proposed an oncology module to facilitate the difficult task of collecting oncology data [10]. The oncology module has a structure for populating chemotherapy episodes and a vocabulary for hematology/oncology [11]. However, a generalizable method of populating chemotherapy episodes has not been proposed, and cases involving the use of data schema are scarce.

Objectives

The main objective of this study was the seamless introduction of the oncology extension into OMOP-CDM by developing an algorithm to automatically identify regimen-level chemotherapy episodes among patients with cancer. To conduct proof-of-concept studies of the availability of the generated chemotherapy episodes, the treatment patterns and trajectories of patients with cancer are presented by additional software. We also identified differences in the onset time and incidence of neutropenia events in patients according to different routine regimens.

Methods

Study Design

This study was composed of two main processes: (1) the development of an algorithm to identify anticancer treatment episodes from the OMOP-CDM database and (2) the analysis of the trends and trajectories in cancer treatment or clinical events based on the algorithm-derived episode records using the visualization software. Furthermore, we performed a pilot study to identify the time of neutropenia onset across various chemotherapy regimens, to validate the scalability of the algorithm. All methods were independently applied to each database and data were collected exclusively as graphical summaries.

Data Sources

We conducted this study using two EHRs of Korean tertiary hospitals and a nationwide claims database from South Korea. The Ajou University School of Medicine (AUSOM) database includes the medical records of 3.14 million patients collected from 1994-2018 [12]. The Kangdong Sacred Heart Hospital (KDH) database includes the medical records of 1.68 million patients collected from 1986-2018. The Health Insurance Review and Assessment Service (HIRA) COVID-19 data set is a nationwide administrative claims database that provides information on reimbursed insurance claims from 2017-2020 for 7590 patients with COVID-19 in South Korea [13]. Each data source was standardized into the OMOP-CDM database (version 5.3). According to the medical history and diagnosis, we identified patients with lung, breast, and colorectal cancer from the EHR databases. From the HIRA COVID-19 data set, we selected patients with COVID-19 and any malignant neoplasm disease as the targets of the descriptive analysis.

Algorithm Development

Workflow of the Chemotherapy Regimen Extraction

Figure 1 shows the entire process used for generating chemotherapy episodes from medication records in harmonized databases. HemOnc is a standard vocabulary adopted by the OMOP-CDM for anticancer agents and chemotherapy regimens derived from the homonymous wiki page, including freely available medical sources for regimens and general information [14]. To convey the semantic regimen protocols in HemOnc to the programmatic algorithm, the section of the protocols

detailing the regimen was collected and converted to parameterized values that list the component drugs as well as appropriate time criteria that reflect the drug schedule. Hence, we developed the Hierarchical Description for Administration of Chemotherapy (HDAC), which is a machine-readable JavaScript Object Notation (JSON) snippet containing the parameterized information of HemOnc. The HDAC contains the variables for the constituent drug of the regimen and the time-range value parameters for the temporal window of the medication schedule.

We developed a Tool for Regimen-level Abstraction of Chemotherapy Episode Records (TRACER) to populate the regimen-level abstracted chemotherapy episodes (Figure 2). The TRACER generated the episodes by leveraging the parameter values of drug conditions or temporal window in HDAC. The chemotherapy episodes included the type of regimen and the number of treatment cycles.

Figure 1. Schematic workflow of the chemotherapy episode extraction. A total of 1506 regimen protocols from HemOnc (a web-based open-source database of cancer chemotherapy regimens) were parameterized as JSON structured data, which is termed HDAC. The JSON file and single drug exposure records in the OMOP-CDM database were instantiated as input data for an algorithm. The TRACER identified chemotherapy episodes by leveraging the parameters from HDAC. The chemotherapy episodes were curated in the episode table, which is an oncology module in the OMOP-CDM. Bev: bevacizumab; CDM: Common Data Model; FOLFIRI: fluorouracil, leucovorin, and irinotecan; FOLFOX: fluorouracil, leucovorin, and oxaliplatin; HDAC: Hierarchical Description for Administration of Chemotherapy; JSON: JavaScript Object Notation; OMOP: Observational Medical Outcomes Partnership; TRACER: Tool for Regimen-level Abstraction of Chemotherapy Episode Record.

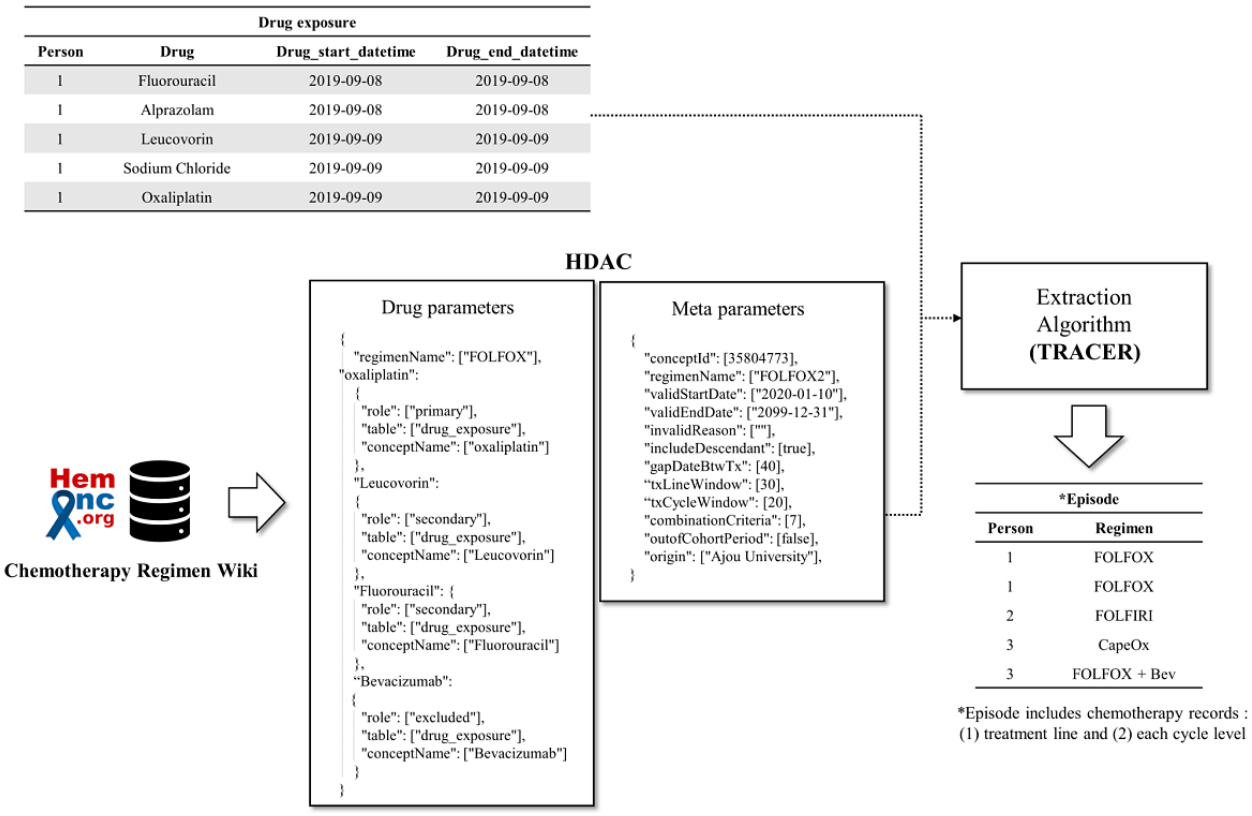
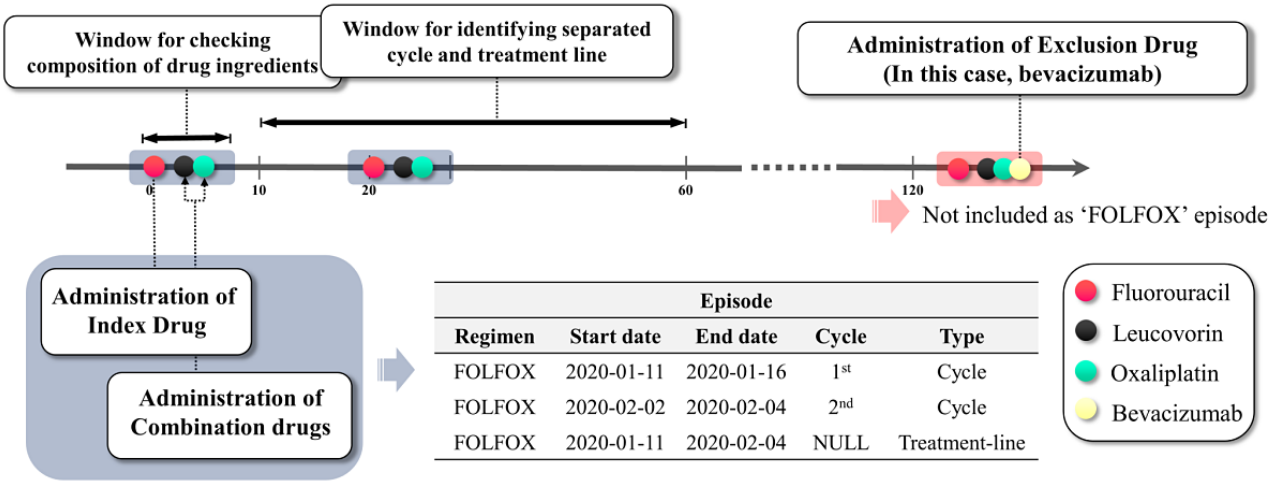


Figure 2. Schematic depiction of algorithm rules in a tool for regimen-level abstraction of chemotherapy episode records. FOLFOX: fluorouracil, leucovorin, and oxaliplatin.



Instantiation of Chemotherapy Regimen Descriptions

The HDAC includes standardized parameters to feed the specifications of regimen protocols into the algorithm. The variables in HDAC are categorized into two types: parameters for drug composition (drug parameters) and meta parameters.

The drug parameter includes the identifiers of drugs (OMOP concept IDs), which link a specific regimen with its respective role. Each drug parameter is granted a role as an index drug, combination drug, or exclusion drug. The index drug in the HDAC is a constituent drug that can be used to identify the first day (day 1) of treatment. A combination drug in the HDAC is a constituent drug (other than the index drug) of the regimen. An exclusion drug is one whose appearance would indicate another regimen. For example, oxaliplatin is the index drug in the fluorouracil, leucovorin, and oxaliplatin (FOLFOX) regimen. Leucovorin and fluorouracil are combination drugs. In this example, bevacizumab is considered as an exclusion drug to distinguish the FOLFOX regimen from the FOLFOX-bevacizumab regimen (Figure 2).

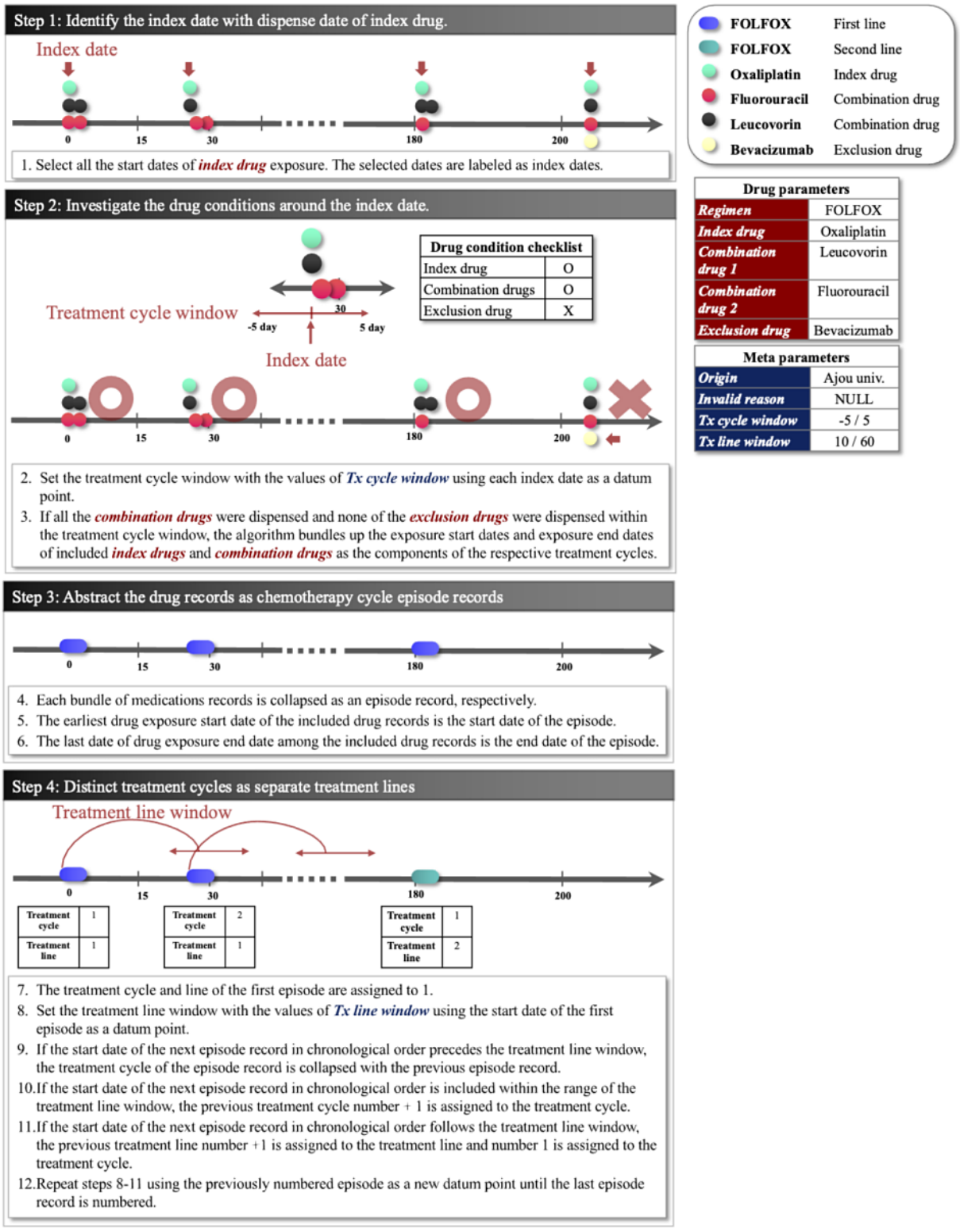
The meta parameter includes the metadata of the HDAC document (eg, origin, valid date, or invalid reason for document) to determine the modifications and define the window range to be adjusted to the algorithm rule. The window identifies a unit of drug records that determines whether the medication record is a part of a particular regimen or distinguishes a boundary for a distinct treatment cycle. The HDAC also stipulates the window for episodes to distinguish the separated treatment line. The concept ID (encoded in the HemOnc vocabulary) of the drug regimen is a primary key for each HDAC snippet. Based on the chemotherapy indications in the HemOnc web database, a total of 1506 indications for chemotherapy protocols were reviewed by an expert and instantiated into the HDAC.

Definitions of the Algorithm

The TRACER sequentially extracts the episodes of regimens included in a list of user settings. The algorithm identifies each treatment cycle episode record and treatment line episode of regimens with the defined rules and parameters in a step-by-step manner (Figure 3). The algorithm consists of the following four steps:

1. Day 1 of the respective treatment cycle (index date) is identified based on the dispense date of the index drug. Each index date is flagged as a datum point for checking the use of other drug ingredients to identify a specific regimen.
2. The prescriptions of the combination drug or exclusion drug are investigated within the period of the predefined window for the cycle in HDAC. If the index drug and all combination drugs were given and none of the exclusion drugs were prescribed in this period, the records of the index drug and combination drug are regarded as a constituent component for a targeting regimen. These records are abstracted as a regimen episode record.
3. The start date of each episode is derived from the start date of the instance of index drug utilization, and the end date of the episode record is derived from the end date of the last index or combination drug utilization. The generated episodes are curated in the episode table of the OMOP-CDM oncology module.
4. The episodes are numbered sequentially in chronological order, provided the interval between the start dates of each episode does not exceed the predefined window in the HDAC. The episode records are collapsed as an identical episode when the interval exceeds the cycle window. The window for distinguishing the treatment line is also defined in the HDAC. The TRACER distinguishes the different treatment lines by changes in regimen type or by episodes beyond the window for the treatment line based on the start date of the previous episode.

Figure 3. Definition of chemotherapy regimen episode extraction algorithm. FOLFIRI: fluorouracil, leucovorin, and irinotecan; FOLFOX: fluorouracil, leucovorin, and oxaliplatin; HDAC: Hierarchical Description for Administration of Chemotherapy.



Algorithm Validation

We reviewed the discharge and progress notes of patients to validate the accuracy of the proposed algorithm. The following regimens were validated: (1) fluorouracil and folinic acid (FULV), (2) FOLFOX, (3) fluorouracil, leucovorin, and irinotecan (FOLFIRI), and (4) capecitabine monotherapy. Among patients with records of algorithm-derived episodes on target regimens, 100 patients were randomly selected for each type of regimen. For this population, we examined the episode records and compared them to clinical notes.

Characterizing the Treatment Patterns and Trajectories

We developed visualization applications for characterizing the treatment patterns for patients with cancer treated with

chemotherapy. Using the tool, we present the relative proportions of the use of the regimens across all chemotherapy treatments from 2008-2018. The distributions of the iterated number of treatment cycles for each patient according to regimen type were portrayed as a heat map with color saturation varying according to the number of patients. The anticancer treatment trajectories were also illustrated for patients who received routine anticancer treatment according to the type of cancer. The trajectories of patients with COVID-19 and cancer were added, to validate the scalability of the tools. The descriptive results of EHR databases included the eight most prevalent regimen types. For the AUSOM database, we added hormone therapies for breast cancer and targeted therapies or immunotherapy for lung cancer in the descriptive analysis.

Timing of Neutropenia Onset Analysis

We also conducted a pilot study that investigated the timing of chemotherapy-induced (febrile) neutropenia (CIN/FN) events from the first chemotherapy episode. Neutropenia is a common adverse event of myelosuppressive chemotherapy. In compliance with National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE; version 5.0) CIN grade 4 (absolute neutrophil count [ANC] $<0.5 \times 10^9$/L) was used to identify severe CIN events. FN events were identified as ANC of $<1.0 \times 10^9$/L with a diagnosis of fever or infection, or any use of granulocyte colony-stimulating factor prophylaxis. A plot showing each neutropenia event as a dot by date of onset was displayed, and in the same plot, a violin plot showed the trends in date of neutropenia onset. To identify the onset time of the neutropenic event, the gap between the date of the first chemotherapy treatment and the date of the first CIN/FN onset was calculated for each patient. As the overall measurement schedule was weekly, the onset dates of neutropenia were categorized in 7-day segments to show the trends of onset dates rather than definite dates. To clarify the effect of a single drug regimen on neutropenia onset, the chemotherapies are limited to the first-line treatment and only the CIN/FN events within 30 days of initiation of chemotherapy were considered. On the day of chemotherapy, the ANC level might be temporarily lowered; therefore, a CIN/FN event during chemotherapy administration was ignored. We also depicted the incidence of CIN/FN events in each cycle by regimen type. The four regimens with the highest frequency of CIN/FN events by type of cancer were included for the incidence plot.

Statistical Analysis

The descriptive analysis was conducted using chemotherapy episodes that were derived from the algorithm. To validate the algorithm, we calculated the proportion of patients who had episodes with identical regimen types as described in clinical notes. The mean absolute error (ie, the mean of the absolute values for the differences between the estimated number of cycles and the actual records in the clinical notes) and the root mean square error (ie, the square root of the mean of the differences between the estimated number of cycles and the actual records in the clinical notes) were also calculated. The overall system was built using R (version 3.5.2; R Foundation for Statistical Computing). The source codes for the algorithm and visualization software have been uploaded to GitHub [15].

Ethics Statement

This study was approved by the institutional review board of Ajou University Hospital of the Republic of Korea (approval number: AJIRB-MED-OBS-20-092) and Kangdong Sacred Heart Hospital of the Republic of Korea (approval number: 2017-03-003). The institutional review board number for the use of HIRA data was AJIRB-MED-EXP-20-087.

Results

Population Characteristics

The TRACER generated a total of 178,360 chemotherapy episodes from the AUSOM database. The episodes consisted of 12 regimen types for colorectal cancer, 24 types for breast cancer (including 6 types of hormone therapy regimen), and 19 types for lung cancer (including 8 types of targeted therapy regimen). The number of patients who were treated with the respective regimens are listed in Multimedia Appendix 1. The characteristics of patients in the AUSOM database are listed (Table 1). Among the 10,353 colorectal, 9546 breast, and 12,671 lung cancer cases, 3151 (30.4%), 5568 (58.3%), and 1593 (12.5%) patients had records of a treatment episode, respectively. The number of patients treated with chemotherapy increased over the years. A total of 69,353 chemotherapy episodes were extracted from the KDH database. The KDH database included 2758 patients with colorectal cancer, 564 (20.4%) of whom had chemotherapy episodes. Among the patients with breast cancer (n=6420) and lung cancer (n=2663) in the KDH database, chemotherapy episodes were identified for 1075 (16.7%) and 642 (24.1%) patients, respectively. The HIRA COVID-19 data set mostly consisted of female patients (60%) [13]. Among 7590 patients with COVID-19, we identified 382 (5%) patients with a primary malignant neoplastic disease.

Table 1. Characteristics of patients with colorectal, breast, and lung cancer in the Ajou University School of Medicine database.

Characteristics of the patients	Patients by type of cancer		
	Colorectal cancer (N=10,353)	Breast cancer (N=9546)	Lung cancer (N=12,671)
Age at index (years), mean (SD)	62 (13.2)	50 (11.4)	64 (12.7)
Sex, n (%)			
Male	6116 (59.1)	62 (0.7)	9166 (72.3)
Female	4237 (40.9)	9484 (99.3)	3505 (27.7)
Number of patients who received chemotherapy by year range, n (%)			
1999–2002	265 (2.6)	646 (6.8)	524 (4.1)
2003–2006	516 (5.0)	829 (8.7)	617 (4.9)
2007–2010	672 (6.5)	965 (10.1)	738 (5.8)
2011–2014	852 (8.2)	1373 (14.4)	775 (6.1)
2015–2018	912 (8.8)	2111 (22.1)	1127 (8.9)
Number of patients who underwent surgery, n (%)	3760 (36.3)	5541 (58.0)	1776 (14.0)
Baseline absolute neutrophil count/μL, mean (SD)	5582 (4403)	3750 (3199)	6517 (5283)
Number of patients who received chemotherapy, n (%)			
First-line treatment	3151 (30.4)	5568 (58.3)	1593 (12.5)
Second-line treatment	1212 (11.7)	4739 (49.6)	888 (7.0)
Third-line treatment	506 (4.8)	4005 (41.9)	521 (4.1)
Fourth-line treatment	234 (2.2)	3573 (37.4)	336 (2.6)

Validation of the Accuracy of TRACER

Table 2 shows the values obtained for the accuracy of the algorithm-derived episode records compared to the clinical notes. The positive predictive value for the type of regimen was over 95% for the FULV regimen, and the chemotherapy type was estimated precisely for the entire episode for FOLFOX, FOLFIRI, and capecitabine monotherapy. The number of treatment cycles was correctly inferred in an average of 85.6% of episode records for validated regimens. The mean difference between the number of cycles in the episode records and the description in the clinical notes was less than one cycle.

Table 2. Validation of chemotherapy episodes compared to chart review.

Treatment regimen	Without information, n	Positive predictive value[a,b] of regimen type, n/N (%)	Accuracy[c] of treatment cycle number, %	Mean absolute error	Root mean square error
FULV[d]	30	67/70 (95)	94	0.1	0.4
FOLFOX[e]	8	92/92 (100)	87	0.3	0.6
FOLFIRI[f]	21	79/79 (100)	89	0.4	1.4
Capecitabine monotherapy	65	35/35 (100)	73	0.7	1.5

[a]For each regimen, 100 cases were randomly sampled and reviewed. The information for chemotherapy was not available in the discharge summary in 30, 8, 21, and 65 cases with FULV, FOLFOX, FOLFIRI, and capecitabine monotherapy, respectively.

[b]Positive predictive value for the matched cases for the type of regimen in the manual comparison of generated episode records with the contents of the clinical notes.

[c]The percentage of matched cases for the number of treatment cycles in the manual comparison of generated episode records with the contents of the clinical notes.

[d]FULV: fluorouracil and leucovorin.

[e]FOLFOX: fluorouracil, leucovorin, and oxaliplatin.

[f]FOLFIRI: fluorouracil, leucovorin, and irinotecan.

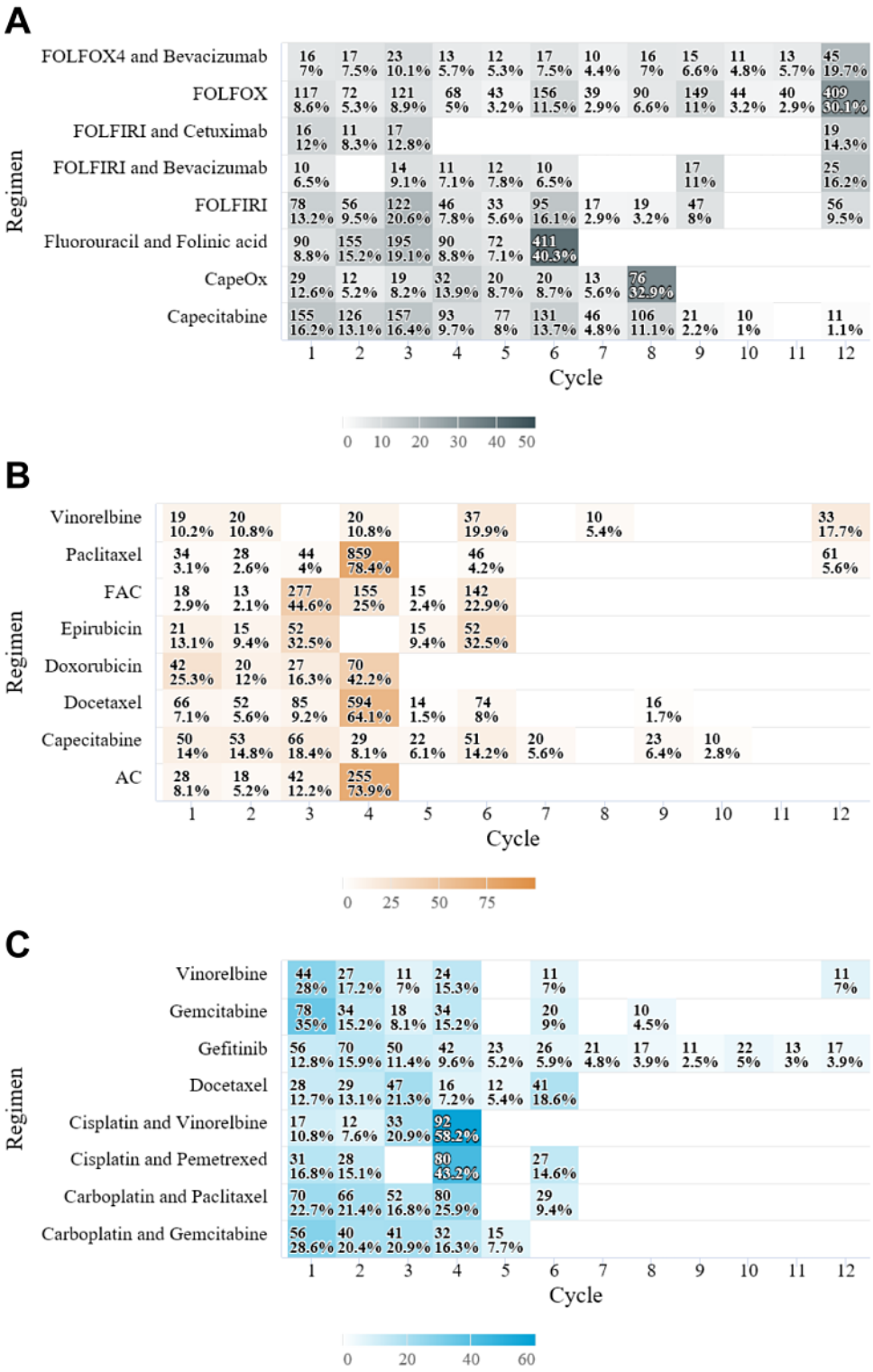
Patterns of Chemotherapy Treatment

The trends in chemotherapy regimen used in the AUSOM database are shown in Multimedia Appendix 2. The number of chemotherapy regimens—including targeted anticancer agents (eg, osimertinib, ceritinib, and crizotinib) and immunotherapies (eg, nivolumab and pembrolizumab)—for lung cancer increased since 2016. On average, 31.2% of patients received the

FOLFOX regimen throughout the years, followed by FULV, which was administered to 20.3% of patients. As of 2012, the proportion of patients receiving chemotherapy with targeted anticancer drugs (eg, bevacizumab or cetuximab) had increased gradually. Tamoxifen use also increased gradually among patients with breast cancer. Multimedia Appendix 3 shows the trends in chemotherapy regimen use in the KDH database. The most frequently used regimen for colorectal cancer in the KDH database was also FOLFOX (27.3% on average), followed by FULV, with an average of 23%. The use of regimens including targeted anticancer drugs for colorectal cancer in the KDH database increased since 2013. From 2013-2016, gefitinib monotherapy was the most frequently used regimen for lung cancer, which was similar to the trends of the AUSOM database. Among the patients with breast cancer in the KDH database, the use of trastuzumab increased sharply since 2014.

The distribution of patients in both EHR databases over the number of iterated chemotherapy cycles is depicted as a heat map. In the heat map of the AUSOM database, the most prevalent number of repeated cycles in the FULV, FOLFOX, and capecitabine and oxaliplatin (CapeOx) regimens for colorectal cancer was consistent with the recommendations of the HemOnc regimen protocol for the adjuvant setting (6, 12, and 8 cycles, respectively; Figure 4). In addition, the most prevalent number of treatment iterations for paclitaxel monotherapy, doxorubicin monotherapy, and cyclophosphamide and doxorubicin (AC) targeting breast cancer was also consistent with the general recommendations (4 cycles in certain cases). “Cisplatin and vinorelbine” and “cisplatin and pemetrexed” for lung cancer were also consistent with the recommendations (4 cycles in certain cases). In the KDH database, the number of patients with colorectal cancer treated with 12 cycles made up the largest proportion among the patients who received the FOLFOX regimen (Multimedia Appendix 4). A total of four regimens—fluorouracil, epirubicin, and cyclophosphamide (FEC); fluorouracil, doxorubicin, and cyclophosphamide (FAC); cyclophosphamide, methotrexate, and fluorouracil (CMF); and Taxotere—for breast cancer were mostly repeated 6 times, and each of the regimen protocols in HemOnc included 6 cycles in certain cases.

Figure 4. Heat map of patient distribution for cycle iteration by regimen type in the Ajou University School of Medicine database. The number of patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer is shown; treatment iteration count is represented by color saturation, with a darker shade representing a higher number of patients. Cells with <10 patients are not reported. AC: doxorubicin and cyclophosphamide; CapeOx: capecitabine and oxaliplatin; FAC: fluorouracil, doxorubicin, and cyclophosphamide; FOLFIRI: fluorouracil, leucovorin, and irinotecan; FOLFOX: fluorouracil, leucovorin, and oxaliplatin.

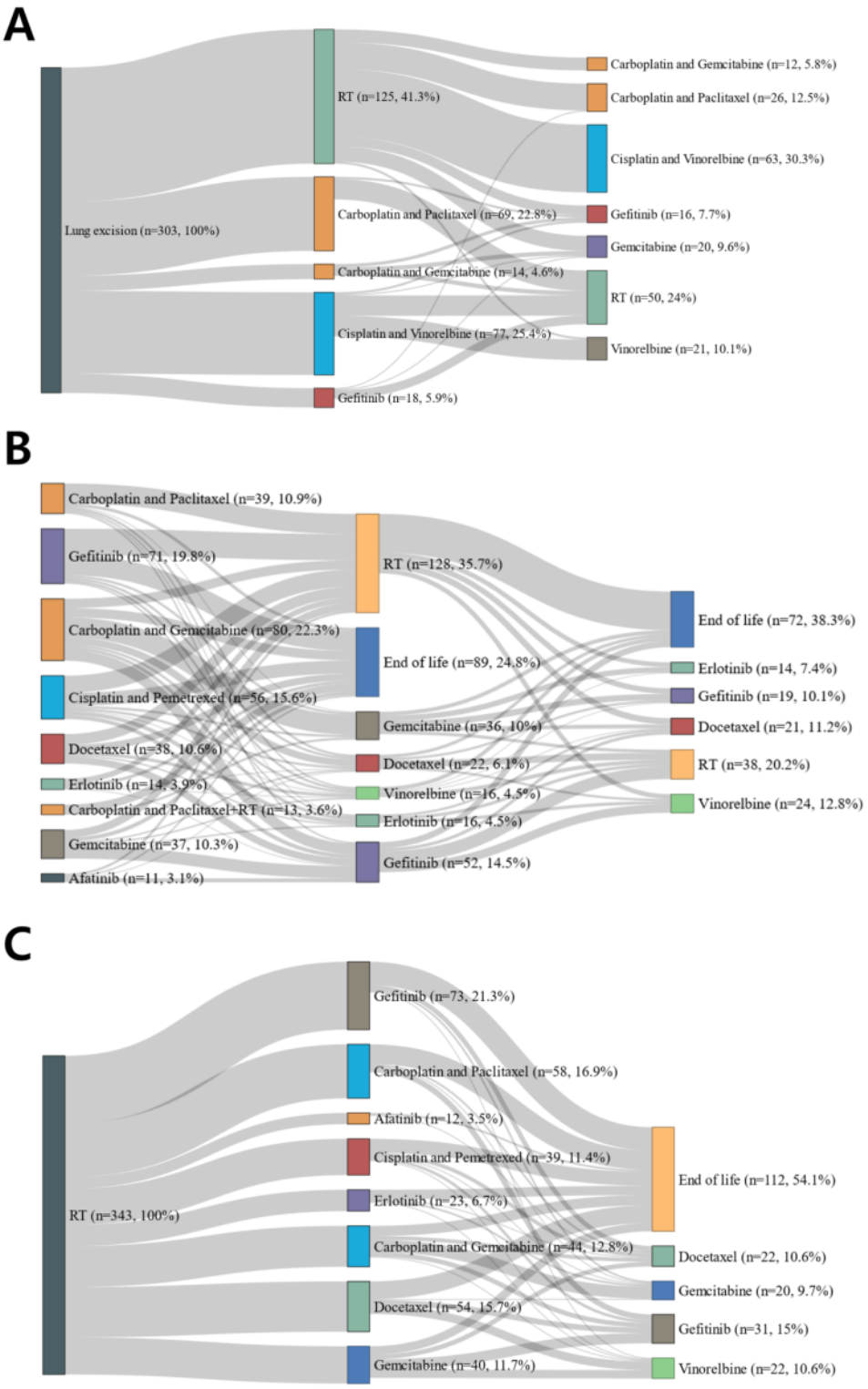


Trajectory of Cancer Treatment

Figure 5 shows the treatment trajectories of patients with lung cancer in the AUSOM database. Among a total of 1120 patients, lung excision–radiation therapy–cisplatin and vinorelbine (n=63, 5.6%) was the most prevalent trajectory. The treatment trajectories of patients with colorectal or breast cancer were displayed, regardless of which first-line treatment was used (Multimedia Appendix 5). Among the treatment trajectories of patients with colorectal cancer or breast cancer that included at least three treatments, colectomy–FOLFOX–FOLFIRI (n=78, 4%) and mastectomy–AC–paclitaxel monotherapy (n=236, 5%)

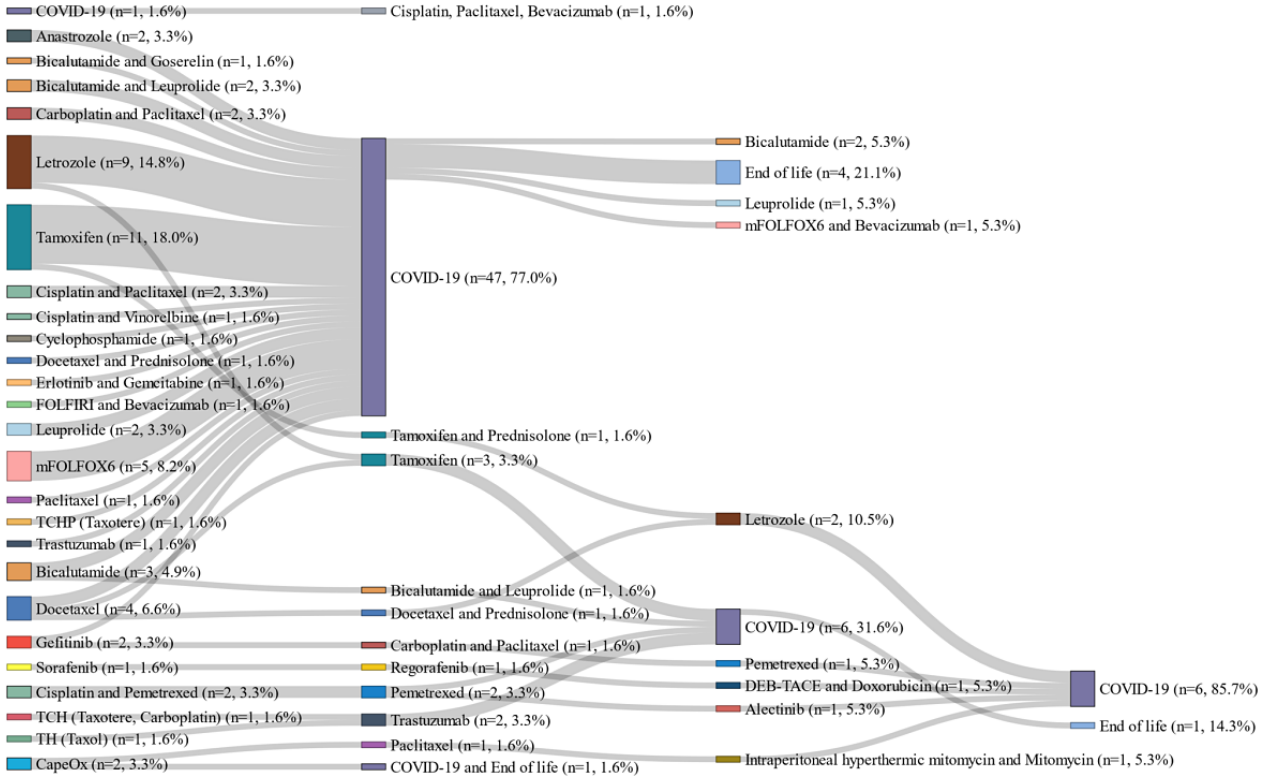
were the trajectories with the highest proportions, respectively. The 10 most frequent trajectories according to cancer type are described in Multimedia Appendix 6.

Figure 5. Anticancer treatment trajectories of patients with lung cancer in the Ajou University School of Medicine database. The treatment trajectories of patients with lung cancer were classified according to the type of first-line treatment: (A) surgery, (B) chemotherapy or chemotherapy with radiation, and (C) radiation therapy. The height of each node represents the population of patients in the corresponding treatment line or therapy. The number of patients who progressed to the next line of treatment is illustrated using gray lines. The chemotherapy regimen changes or the transition between types of treatment were regarded as a treatment line transition. The percentage on the label covers the proportion of the number of patients to all patients in the identical line of trajectory. As the large number of nodes hinders the purpose of the visualizations within a graphical summary, the nodes are truncated at the third node. For the same reason, the nodes for patient count under 10 were removed. AUSOM: Ajou University School of Medicine; RT: radiation therapy.



The treatment trajectories of patients in the KDH database are illustrated in Multimedia Appendix 7. Among the patients with colorectal cancer, FOLFOX was the most frequently used first-line (n=52, 17.9%) and second-line (n=84, 28.9%) regimen in the trajectory. For breast cancer, Taxotere was the most frequent first-line chemotherapy before mastectomy (n=58, 17.3%). Gefitinib was the most widely used first-line regimen among patients with lung cancer (n=30, 10.5%). Figure 6 shows the treatment trajectories of patients with malignant neoplasm who also had COVID-19. Of the 7590 patients nationwide with a diagnosis of COVID-19, we identified 382 patients with a history of cancer. Among them, a total of 62 patients had an episode of chemotherapy. Most of the patients received only one line of treatment between 2017-2020, before COVID-19 infection (n=47). The trajectory included 6 patients with a node of end of life.

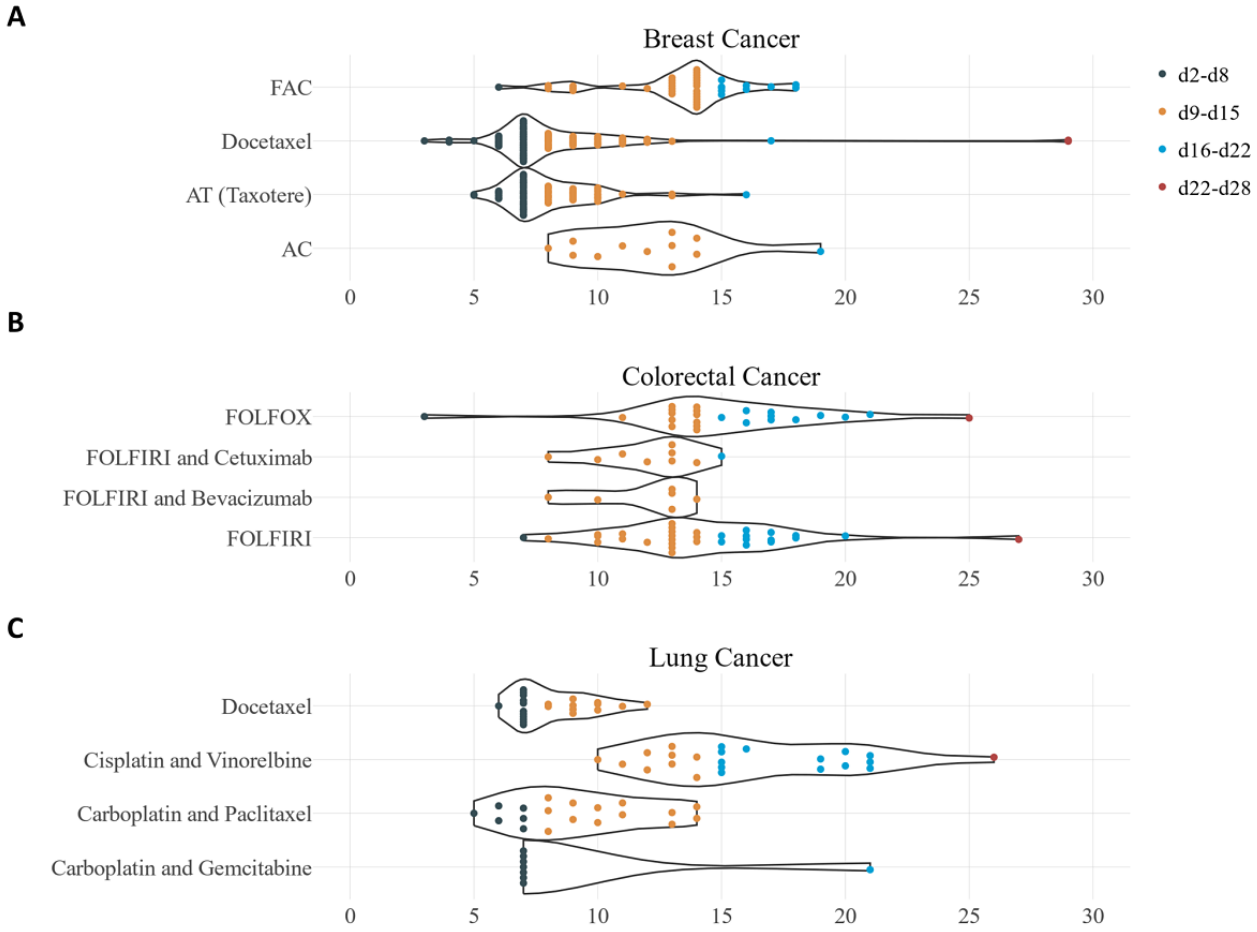
Figure 6. Anticancer treatment trajectories of patients with COVID-19. Sankey plot of the treatment trajectories of patients with COVID-19, including episodes of anticancer chemotherapy between 2017 and 2020. Each node represents the chemotherapy regimen used for cancer treatment. The percentage on the label covers the proportion of the patient number in each node in the same trajectory phase. As the large number of nodes hinders the purpose of the visualizations within a graphical summary, the nodes are truncated at the fourth node. For the same reason, the nodes for patient count <5 were removed.



Timing of Chemotherapy-Induced Neutropenia

Figure 7 shows the time of onset of the CIN/FN event for each patient in the AUSOM database. The episodes of neutropenia among patients with colorectal cancer were clustered between days 9 and 15. Compared with the regimens used for colorectal cancer, the neutropenia events that were recorded after docetaxel monotherapy and Taxotere treatment for breast cancer and after carboplatin and gemcitabine for lung cancer generally began one week earlier (days 2-8). We illustrated the incidence of neutropenia events in each cycle of treatment in Multimedia Appendix 8. Regardless of the cancer type, the incidence of CIN/FN events was high during the first cycle, with the exception of the FOLFOX regimen for colorectal cancer and the carboplatin and paclitaxel regimen for lung cancer. Finally, neutropenia occurred more frequently during the Taxotere regimen for breast cancer (75.3%), which includes doxorubicin and docetaxel as constituent drugs.

Figure 7. Trends in neutropenia onset time according to regimen. Time of onset of chemotherapy-induced (febrile) neutropenia event after the first exposure to chemotherapy among patients with (A) breast cancer, (B) colorectal cancer, and (C) lung cancer at the Ajou University School of Medicine. Each dot represents the neutropenia event of a distinct patient. The events are categorized in a 7-day range. The violin plot represents the trends in the frequency on each day from chemotherapy exposure. AC: doxorubicin and cyclophosphamide; FAC: fluorouracil, doxorubicin, and cyclophosphamide; FOLFIRI: fluorouracil, leucovorin, and irinotecan; FOLFOX: fluorouracil, leucovorin, and oxaliplatin.



Discussion

Overview

This study described a system for analyzing the treatment patterns and trajectories of patients with cancer based on the oncology extension model in the OMOP-CDM. The proposed algorithm (TRACER) for extracting chemotherapy episodes at the regimen level effectively generated the treatment episodes for patients with cancer. This approach illustrates how laborious manual curation can be replaced with an automatic extraction system. The obtained episodes were validated by reviewing clinical notes, which revealed that the type of regimen or the number of treatment cycles were estimated with high accuracy. We also demonstrated the usefulness of the proposed system by performing a pilot study investigating the onset time of CIN/FN across various chemotherapy regimens.

Principal Findings

Comprehensive clinical information, including longitudinal treatment sequences and the various clinical outcomes of patients with cancer, is not available in nationwide cancer registries, such as the Surveillance, Epidemiology, and End Results Program [16-18] or the Korea Central Cancer Registry [19,20]. Large-scale real-world data derived from EHRs [21] and administrative claims data [17] of a standardized data network can support the timely assessment of the characterization and quality of routine clinical practice and active pharmacovigilance across institutions or countries.

The unexpectedly rapid spread of COVID-19 revealed an urgent unmet need for the timely retrieval of detailed data for patients with cancer, to provide relevant evidence for the management of patients with cancer during the pandemic period [22]. Although conventional cancer registries have failed to provide these data to researchers, the secondary use of EHRs and claims databases can promptly provide valuable insights into the impact of a novel infection on patients with cancer [13]. The TRACER was able to generate records to describe the trajectory of cancer treatment and death of patients with COVID-19, which may be helpful for identifying the relationship between cancer treatment and a fatal case of COVID-19.

We demonstrated how electronically captured data elements can support clinical research using longitudinal detailed clinical data. FN is one of the most common oncologic emergencies [23] and is associated with considerable morbidity and mortality [24]. Although it is well known that the risk of CIN/FN is highest during the first cycle of chemotherapy for solid tumors

or lymphoma [25,26], the exact time of occurrence of CIN/FN in various regimens is largely unknown. We found that CIN/FN events were more frequent in the first cycle of chemotherapy, and that regimens that included docetaxel or doxorubicin were followed by a greater number of CIN/FN events, which is compatible with reported findings [27]. CIN/FN events usually occurred relatively early (days 2-8) in patients who received regimens including docetaxel or carboplatin compared with those treated with other regimens (days 7-13).

Limitations

This study had several limitations. First, only four of the regimens were validated through manual review; thus, it was not clear whether episodes for the other types of regimen can also be precisely estimated. Nevertheless, the patterns of treatment cycle repetition showed that the extracted records were concordant with the standard protocols of the regimens, suggesting that the algorithm could correctly interpret the variable regimens. A second limitation was that the relatively low rate of ascertained treatment episodes (maximum of 58% among patients with breast cancer in the AUSOM database) suggests that treatment episodes may have been missed, although many patients with cancer are treated with surgery or radiation alone and would have appropriately not been captured. This may be due to the fact that the algorithm extracts only the regimens included in the HemOnc vocabulary. The flexible structure of HDAC, which allows the addition of user-defined rules for specific regimens, has the potential to mitigate the missing rate of treatment episodes for a particular study using a fine-tuned algorithm.

Conclusions

We developed a technique to generate episodes for chemotherapies included in the oncology module of the OMOP-CDM and to analyze treatment patterns in patients with cancer. We demonstrated that the proposed process is reproducible and scalable across a distributed data network. Our findings suggest that a generalizable strategy of characterizing treatment trajectories from harmonized observational databases can promptly determine the characteristics of clinical events, thus enabling the generation of real-world evidence for abrupt pandemic crises. Further research is required to generate statistical evidence for clinical outcomes between regimen types.

Acknowledgments

SCY, HJ, and RWP contributed to the study design. All authors contributed to the writing and final approval of this manuscript. This work was supported by the Bio Industrial Strategic Technology Development Program (20001234) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI16C0992). JLW was supported by the National Cancer Institute (grant number CA231840).

Conflicts of Interest

JW has equity in HemOnc.org LLC; this equity has no financial value. JW is the chief software architect of the HemOnc ontology; this role is uncompensated. The development of HemOnc is partially supported by NIH grants U24 CA194215; U01 CA231840; and U24 CA248010. The funder had no role in the development or conduct of the study.

Multimedia Appendix 1
Chemotherapy episodes from the Ajou University School of Medicine database. The 10 most frequently used chemotherapy regimens for colorectal cancer, breast cancer, and lung cancer are listed.
[DOCX File , 23 KB - medinform_v9i4e25035_app1.docx]

Multimedia Appendix 2
Trends of chemotherapy regimen use within the Ajou University School of Medicine database. The proportions of chemotherapy regimen uses for patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer by year from 2008-2018 in the Ajou University School of Medicine database are shown.
[DOCX File , 492 KB - medinform_v9i4e25035_app2.docx]

Multimedia Appendix 3
Trends of chemotherapy regimen use within the Kangdong Sacred Heart Hospital database. The proportions of chemotherapy regimen uses for patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer by year from 2008-2018 in the Kangdong Sacred Heart Hospital database are shown.
[DOCX File , 167 KB - medinform_v9i4e25035_app3.docx]

Multimedia Appendix 4
Heat map of patient distribution for cycle iteration by regimen type in the Kangdong Sacred Heart Hospital database. The number of patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer is shown; treatment iteration counts are represented by a color saturation difference.

[DOCX File , 837 KB - medinform_v9i4e25035_app4.docx]

Multimedia Appendix 5
Trends of chemotherapy regimen use from the Kangdong Sacred Heart Hospital database. The proportions of chemotherapy regimen uses for patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer by year from 2008-2018 in the Kangdong Sacred Heart Hospital database are shown.
[DOCX File , 561 KB - medinform_v9i4e25035_app5.docx]

Multimedia Appendix 6
The list of treatment trajectories of patients with cancer from the Ajou University School of Medicine database.
[DOCX File , 21 KB - medinform_v9i4e25035_app6.docx]

Multimedia Appendix 7
Anticancer treatment trajectories for patients with cancer in the Kangdong Sacred Heart Hospital database. The treatment trajectories of patients with (A) colorectal cancer, (B) breast cancer, and (C) lung cancer in the Kangdong Sacred Heart Hospital database are shown.
[DOCX File , 605 KB - medinform_v9i4e25035_app7.docx]

Multimedia Appendix 8
Incidence of neutropenia by treatment cycle. The histogram of incidence of the first neutropenia event by cycle and regimen for (A) colorectal cancer, (B) lung cancer, and (C) breast cancer.
[DOCX File , 165 KB - medinform_v9i4e25035_app8.docx]

References

1. Khozin S, Blumenthal GM, Pazdur R. Real-world Data for Clinical Evidence Generation in Oncology. J Natl Cancer Inst 2017 Nov 01;109(11). [doi: 10.1093/jnci/djx187] [Medline: 29059439]
2. Bendell JC, Bekaii-Saab TS, Cohn AL, Hurwitz HI, Kozloff M, Tezcan H, et al. Treatment patterns and clinical outcomes in patients with metastatic colorectal cancer initially treated with FOLFOX-bevacizumab or FOLFIRI-bevacizumab: results from ARIES, a bevacizumab observational cohort study. Oncologist 2012;17(12):1486-1495 [FREE Full text] [doi: 10.1634/theoncologist.2012-0190] [Medline: 23015662]
3. Dasari A, Bergsland EK, Benson AB, Cai B, Huynh L, Totev T, et al. Treatment Patterns and Clinical Outcomes in Advanced Lung Neuroendocrine Tumors in Real-World Settings: A Multicenter Retrospective Chart Review Study. Oncologist 2019 Aug;24(8):1066-1075 [FREE Full text] [doi: 10.1634/theoncologist.2018-0520] [Medline: 30610008]
4. Nadler E, Espirito JL, Pavilack M, Boyd M, Vergara-Silva A, Fernandes A. Treatment Patterns and Clinical Outcomes Among Metastatic Non-Small-Cell Lung Cancer Patients Treated in the Community Practice Setting. Clin Lung Cancer 2018 Jul;19(4):360-370 [FREE Full text] [doi: 10.1016/j.cllc.2018.02.002] [Medline: 29576407]
5. Bikov KA, Mullins CD, Seal B, Onukwugha E, Hanna N. Algorithm for identifying chemotherapy/biological regimens for metastatic colon cancer in SEER-Medicare. Med Care 2015 Aug;53(8):e58-e64. [doi: 10.1097/MLR.0b013e31828fad9f] [Medline: 23552436]
6. Carroll NM, Burniece KM, Holzman J, McQuillan DB, Plata A, Ritzwoller DP. Algorithm to Identify Systemic Cancer Therapy Treatment Using Structured Electronic Data. JCO Clin Cancer Inform 2017 Nov;1:1-9 [FREE Full text] [doi: 10.1200/CCI.17.00002] [Medline: 30657379]
7. Prodel M, Lamarsalle L, Augusto V. ATLAS: a robust algorithm for temporal sequence alignment of treatment lines using claim databases. 2019 Presented at: 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); July 9-11, 2019; Siena, Italy p. 1-8. [doi: 10.1109/cibcb.2019.8791467]
8. Weymann D, Costa S, Regier DA. Validation of a Cyclic Algorithm to Proxy Number of Lines of Systemic Cancer Therapy Using Administrative Data. JCO Clin Cancer Inform 2019 Aug;3:1-10 [FREE Full text] [doi: 10.1200/CCI.19.00022] [Medline: 31365273]
9. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
10. Belenkaya R, Gurley M, Dymshyts D, Araujo S, Williams A, Chen R, et al. Standardized Observational Cancer Research Using the OMOP CDM Oncology Module. Stud Health Technol Inform 2019 Aug 21;264:1831-1832. [doi: 10.3233/SHTI190670] [Medline: 31438365]
11. Warner JL, Dymshyts D, Reich CG, Gurley MJ, Hochheiser H, Moldwin ZH, et al. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. J Biomed Inform 2019 Aug;96:103239 [FREE Full text] [doi: 10.1016/j.jbi.2019.103239] [Medline: 31238109]

12. Yoon D, Ahn EK, Park MY, Cho SY, Ryan P, Schuemie MJ, et al. Conversion and Data Quality Assessment of Electronic Health Record Data at a Korean Tertiary Teaching Hospital to a Common Data Model for Distributed Network Research. Healthc Inform Res 2016 Jan;22(1):54-58 [FREE Full text] [doi: 10.4258/hir.2016.22.1.54] [Medline: 26893951]
13. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. medRxiv. Preprint published online on April 25, 2020 [FREE Full text] [doi: 10.1101/2020.04.22.20074336] [Medline: 32511443]
14. Warner JL, Cowan AJ, Hall AC, Yang PC. HemOnc.org: A Collaborative Online Knowledge Platform for Oncology Professionals. J Oncol Pract 2015 May;11(3):e336-e350 [FREE Full text] [doi: 10.1200/JOP.2014.001511] [Medline: 25736385]
15. Jeon HK, You SC. CancerTxPathway. GitHub. 2020. URL: https://github.com/ABMI/CancerTxPathway [accessed 2020-10-15]
16. Bach PB, Guadagnoli E, Schrag D, Schussler N, Warren JL. Patient demographic and socioeconomic characteristics in the SEER-Medicare database applications and limitations. Med Care 2002 Aug;40(8 Suppl):IV-19. [doi: 10.1097/00005650-200208001-00003] [Medline: 12187164]
17. Enewold L, Parsons H, Zhao L, Bott D, Rivera DR, Barrett MJ, et al. Updated Overview of the SEER-Medicare Data: Enhanced Content and Applications. J Natl Cancer Inst Monogr 2020 May 01;2020(55):3-13. [doi: 10.1093/jncimonographs/lgz029] [Medline: 32412076]
18. Warren JL, Klabunde CN, Schrag D, Bach PB, Riley GF. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. Med Care 2002 Aug;40(8 Suppl):IV-I3. [doi: 10.1097/01.MLR.0000020942.47004.03] [Medline: 12187163]
19. Jung K, Won Y, Kong H, Lee ES. Cancer Statistics in Korea: Incidence, Mortality, Survival, and Prevalence in 2016. Cancer Res Treat 2019 Apr;51(2):417-430 [FREE Full text] [doi: 10.4143/crt.2019.138] [Medline: 30913865]
20. Shin H, Won Y, Jung K, Kong H, Yim S, Lee J, Members of the Regional Cancer Registries. Nationwide cancer incidence in Korea, 1999~2001; first result using the national cancer incidence database. Cancer Res Treat 2005 Dec;37(6):325-331 [FREE Full text] [doi: 10.4143/crt.2005.37.6.325] [Medline: 19956367]
21. Ma X, Long L, Moon S, Adamson B, Baxi S. Comparison of Population Characteristics in Real-World Clinical Oncology Databases in the US: Flatiron Health, SEER, and NPCR. MedRxiv. Preprint published online on May 30, 2020. [doi: 10.1101/2020.03.16.20037143]
22. Kuderer NM, Choueiri TK, Shah DP, Shyr Y, Rubinstein SM, Rivera DR, COVID-19 Cancer Consortium. Clinical impact of COVID-19 on patients with cancer (CCC19): a cohort study. Lancet 2020 Jun 20;395(10241):1907-1918 [FREE Full text] [doi: 10.1016/S0140-6736(20)31187-9] [Medline: 32473681]
23. Lewis MA, Hendrickson AW, Moynihan TJ. Oncologic emergencies: Pathophysiology, presentation, diagnosis, and treatment. CA Cancer J Clin 2011;61(5):287-314 [FREE Full text] [doi: 10.3322/caac.20124] [Medline: 21858793]
24. Kuderer NM, Dale DC, Crawford J, Cosler LE, Lyman GH. Mortality, morbidity, and cost associated with febrile neutropenia in adult cancer patients. Cancer 2006 May 15;106(10):2258-2266 [FREE Full text] [doi: 10.1002/cncr.21847] [Medline: 16575919]
25. Culakova E, Thota R, Poniewierski MS, Kuderer NM, Wogu AF, Dale DC, et al. Patterns of chemotherapy-associated toxicity and supportive care in US oncology practice: a nationwide prospective cohort study. Cancer Med 2014 Apr;3(2):434-444 [FREE Full text] [doi: 10.1002/cam4.200] [Medline: 24706592]
26. Crawford J, Dale DC, Kuderer NM, Culakova E, Poniewierski MS, Wolff D, et al. Risk and timing of neutropenic events in adult cancer patients receiving chemotherapy: the results of a prospective nationwide study of oncology practice. J Natl Compr Canc Netw 2008 Feb;6(2):109-118. [doi: 10.6004/jnccn.2008.0012] [Medline: 18319047]
27. Lalami Y, Paesmans M, Muanza F, Barette M, Plehiers B, Dubreucq L, et al. Can we predict the duration of chemotherapy-induced neutropenia in febrile neutropenic patients, focusing on regimen-specific risk factors? A retrospective analysis. Ann Oncol 2006 Mar;17(3):507-514 [FREE Full text] [doi: 10.1093/annonc/mdj092] [Medline: 16322116]

Abbreviations

AC: cyclophosphamide and doxorubicin
ANC: absolute neutrophil count
AUSOM: Ajou University School of Medicine
CapeOx: capecitabine and oxaliplatin
CDM: common data model
CIN/FN: chemotherapy-induced (febrile) neutropenia
CMF: cyclophosphamide, methotrexate, and fluorouracil
CRCAE: Common Terminology Criteria for Adverse Events
EHR: electronic health record
FAC: fluorouracil, doxorubicin, and cyclophosphamide
FEC: fluorouracil, epirubicin, and cyclophosphamide

FOLFIRI: fluorouracil, leucovorin, and irinotecan
FOLFOX: fluorouracil, leucovorin, and oxaliplatin
FULV: fluorouracil and folinic acid
HDAC: Hierarchical Description for Administration of Chemotherapy
HIRA: Health Insurance Review and Assessment Service
JSON: JavaScript Object Notation
KDH: Kangdong Sacred Heart Hospital
OHDSI: Observational Health Data Sciences and Informatics
OMOP: Observational Medical Outcomes Partnership
TRACER: Tool for Regimen-level Abstraction of Chemotherapy Episode Records

Edited by G Eysenbach; submitted 15.10.20; peer-reviewed by J Banda, L Rusu; comments to author 07.11.20; revised version received 12.11.20; accepted 20.01.21; published 06.04.21.

Please cite as:
Jeon H, You SC, Kang SY, Seo SI, Warner JL, Belenkaya R, Park RW
Characterizing the Anticancer Treatment Trajectory and Pattern in Patients Receiving Chemotherapy for Cancer Using Harmonized Observational Databases: Retrospective Study
JMIR Med Inform 2021;9(4):e25035
URL: https://medinform.jmir.org/2021/4/e25035
doi:10.2196/25035
PMID:33720842

Original Paper

Factors Affecting General Practitioners' Readiness to Accept and Use an Electronic Health Record System in the Republic of North Macedonia: A National Survey of General Practitioners

Tomi Dimitrovski[1,2*], PhD; Peter A Bath[3,4*], PhD; Panayiotis Ketikidis[1,2*], PhD; Lambros Lazuras[5*], PhD

[1]CITY College, University of York Europe Campus, Thessaloniki, Greece

[2]South-East European Research Centre, Thessaloniki, Greece

[3]Information School, University of Sheffield, Sheffield, United Kingdom

[4]School of Health and Related Research, University of Sheffield, Sheffield, United Kingdom

[5]Department of Psychology, Sociology & Politics, Sheffield Hallam University, Sheffield, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Tomi Dimitrovski, PhD
CITY College
University of York Europe Campus
24 Proxenou Koromila Street
Thessaloniki, 54622
Greece
Phone: 30 6979222130
Email: tdimitrovski@citycollege.sheffield.eu

Abstract

Background: Electronic health records (EHRs) represent an important aspect of digital health care, and to promote their use further, we need to better understand the drivers of their acceptance among health care professionals. EHRs are not simple computer applications; they should be considered as a highly integrated set of systems. Technology acceptance theories can be used to better understand users' intentions to use EHRs. It is recommended to assess factors that determine the future acceptance of a system before it is implemented.

Objective: This study uses a modified version of the Unified Theory of Acceptance and Use of Technology with the aim of examining the factors associated with intentions to use an EHR application among general practitioners (GPs) in the Republic of North Macedonia, a country that has been underrepresented in extant literature. More specifically, this study aims to assess the role of technology acceptance predictors such as performance expectancy, effort expectancy, social influence, facilitating conditions, job relevance, descriptive norms, and satisfaction with existing eHealth systems already implemented in the country.

Methods: A web-based invitation was sent to 1174 GPs, of whom 458 completed the questionnaire (response rate=40.2%). The research instrument assessed performance expectancy, effort expectancy, facilitating conditions, and social influence in relation to the GPs' intentions to use future EHR systems. Job relevance, descriptive norms, satisfaction with currently used eHealth systems in the country, and computer/internet use were also measured.

Results: Hierarchical linear regression analysis showed that effort expectancy, descriptive norms, social influence, facilitating conditions, and job relevance were significantly associated with intentions to use the future EHR system, but performance expectance was not. Multiple mediation modeling analyses further showed that social influence (z=2.64; P<.001), facilitating conditions (z=4.54; P<.001), descriptive norms (z=4.91; P<.001), and effort expectancy (z=5.81; P=.008) mediated the association between job relevance and intentions. Finally, moderated regression analysis showed that the association between social influence and usage intention was significantly moderated (P=.02) by experience ($B_{experience \times social\ influence}$=.005; 95% CI 0.001 to 0.010; β=.080). In addition, the association between social influence and intentions was significantly moderated (P=.02) by age ($B_{age \times social\ influence}$=.005; 95% CI 0.001 to 0.010; β=.077).

Conclusions: Expectations of less effort in using EHRs and perceptions on supportive infrastructures for enabling EHR use were significantly associated with the greater acceptance of EHRs among GPs. Social norms were also associated with intentions,

even more so among older GPs and those with less work experience. The theoretical and practical implications of these findings are also discussed.

(JMIR Med Inform 2021;9(4):e21109) doi:10.2196/21109

KEYWORDS

general practitioner; eHealth; technology acceptance; electronic health record

Introduction

Background

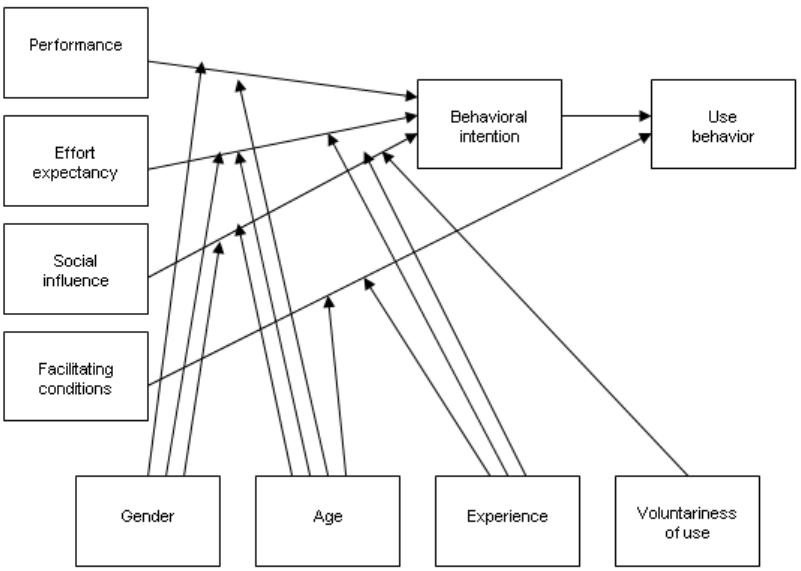
The development and implementation of eHealth and digital technologies in health care have become widespread in recent decades. However, there have been numerous failures in eHealth systems because of the lack of adoption and use of these technologies and systems by health care professionals and other staff in health care systems [1,2]. The underutilization of digital technology in health care settings is evident, although the reasons for this are unclear. The low acceptance of new technologies in health care settings remains to be a challenge for health service management and researchers [1,3,4]. Therefore, it is important to gain a better understanding of the processes underlying health care professionals' acceptance of novel health care technologies and systems [5-7].

Electronic health record (EHR) systems are an essential part of information and communication technologies (ICTs) within health care settings and organizations. In primary health care, EHR systems have been developed to support the storage, retrieval, and use of patient data over the life course of a patient by general practitioners (GPs) and other health care professionals in primary care.

Technology Acceptance

Different theories have been developed to assess the factors that influence the adoption and use of ICTs in health care, including the Unified Theory of Acceptance and Use of Technology (UTAUT; Figure 1) [8], which seeks to understand the effect of various factors on users' intentions to use a new system, as well as their actual use of the system. The 4 basic technology acceptance constructs within the UTAUT are performance expectancy, effort expectancy, social influence, and facilitating conditions. Performance expectancy assesses an individual's anticipation of improved performance resulting from the use of new technologies. Effort expectancy represents the end users' perceptions of the ease of using new ICTs (ie, how much effort will be required by them to use the new system). Social influence measures the subjective social norms of end users and represents referent others' endorsement of using the technology in question and the perceived prevalence of the utilization of the technology in referent groups. Facilitating conditions represent the degree to which end users perceive that there will be organizational and technical support for the efficient and easy use of the technology [8]. The original UTAUT model has 4 potential moderators: gender, age, experience, and voluntariness. This means that the association between the UTAUT constructs and usage intentions may be stronger or weaker, depending on the values of the moderator constructs (eg, the association between performance expectancy and intentions to use the technology may be stronger among individuals with more vs less experience in using the technology) [8].

Figure 1. The Unified Theory of Acceptance and Use of Technology.



Although the application of early technology acceptance models in health care settings started in the late 1990s [9,10], there is still limited empirical research on technology acceptance in EHR systems. Research on technology acceptance in health care has suggested that performance expectancy is the strongest and most important predictor of intentions to use EHR systems [8,11-14]. Effort expectancy has been shown to be a significant predictor of intentions to use EHR systems [13-17] in a smaller number of studies, and social influence and facilitating conditions have rarely been investigated [14]. A number of additional technology acceptance constructs have been applied in several studies in health care settings, including health information technology experience [7,18], computer knowledge [19], job relevance [20], and the self-assessment of computer use at home [19]. For this research, these constructs can be considered in their original or modified forms. The UTAUT model was applied in mandatory health care settings (where EHR use is compulsory) in the relevant literature [13,14,20].

This Study

This study is a part of a PhD thesis and is published for the first time in a journal. A national EHR system has been proposed for the Republic of North Macedonia, and the aim of this study is to examine the factors that influence the adoption of such a system among GPs within the country. All GPs in the country worked in private settings, but they had active contracts with the National Health Insurance Fund and were obliged to follow the work instructions proposed by the fund. The proposed EHR system was not implemented in the country when this research was conducted. The technology acceptance assessment was conducted before the implementation of the EHR system in the country with the aim of identifying the factors that determine intentions for future use. However, the "Health Smart Card" system (a smart card access to basic patient personal data and health insurance) and the "My term system" (a web-based scheduling system) were implemented in the country at the time when this research was conducted.

The main objective of this research is to assess the readiness of GPs in the country for the future acceptance of EHR systems. Other objectives are to address the role of the basic predictors of the original UTAUT model on EHR use; to assess the effect of other technology acceptance predictors such as job relevance, descriptive norm, and satisfaction (with existing health ICT systems already implemented in the country); and to identify the moderating effect of basic moderation variables such as age, gender, and previous work experience.

Adding new technology acceptance constructs to the basic UTAUT model was an opportunity to develop a better understanding of the factors influencing the use of ICTs in a large sample of GPs within a country. However, some technology acceptance constructs, such as descriptive norm [21], computer use, internet use, and use of other technology [22-26] were derived and modified from the referent literature studies on technology acceptance and were identified as useful for this research. Descriptive norms can be regarded as a measure of the potential use of EHRs by colleagues.

The following hypotheses were developed:

- H1: the original UTAUT constructs (ie, performance expectancy, effort expectancy, social influence, and facilitating conditions) will be associated with intentions to use the EHR system in the future.
- H2: other technology acceptance constructs—job relevance, satisfaction, and the use of other technology—will be indirectly associated with intentions to use the EHR system in the future through the effects of performance expectancy.
- H3: the association between the basic UTAUT constructs and intentions to use the future EHR system will be moderated by age, gender, and previous work experience (moderation effect according to the UTAUT model).
- H4: descriptive norms will be significantly associated with intentions to use the EHR system in the future, over and above the effects of other predictor constructs.

The assessment of the hypotheses identifies the effects of technology acceptance variables on user intentions. Therefore, this research aims to establish the most important technology acceptance predictors for future EHR systems among GPs in the country.

Methods

Recruitment

The target population was the GPs in the country; all GPs who had contracts with the National Health Insurance Fund were included in the study. Participants' email addresses were provided by the National Health Insurance Fund List. According to the list, there were 1631 active GPs in the country at the time of the study, with 1174 active email addresses of GPs registered in the list. A web-based survey was created on the SharePoint (TM) platform, and an invitation email was sent to all email addresses. General information on the future EHR system is included in a short introduction to the survey. The email was sent on July 1, 2014, followed by 2 reminder emails on July 15, 2014, and August 1, 2014. However, 35 emails were returned, as they did not reach valid email addresses.

Research Instrument

The original UTAUT model was modified with other technology acceptance extensions for this study. The following technology acceptance items were added to the questionnaire: job relevance [11], descriptive norm (ie, estimated prevalence of EHR use by colleagues in the future) [21,22], current use of other technology for professional or leisure purposes [23], and satisfaction with existing eHealth systems that are currently used in the country. A (user) satisfaction item was developed to assess the GPs' satisfaction with the currently used ICT systems in health care in the country (the "Health Smart Card" system and the "My term system"). The purpose of including this item was to assess the association of user satisfaction with existing health care ICT systems with the intention of using the future EHR. Job relevance was added to the current research model, as its effectiveness was established in a previous study conducted by researchers [22].

Performance expectancy [8,15] was measured by using 5 questions for assessing aspects of participants' beliefs about the usefulness of future EHR systems. Effort expectancy

[8,11,12,15] was measured by using 8 items for assessing aspects of the ease of use of the future EHR system. Facilitating conditions [8] were measured with 4 items for assessing the degree to which participants believed that organizational infrastructure would support their use of the future EHR system. Social influence [8,12,15] was measured with the mean scores of 3 items for assessing how a participant perceived other colleagues' beliefs about whether they should use the future EHR system. The descriptive norm [21] variable was measured with a single item that asked participants to estimate how many of their colleagues would use the proposed EHR system if it was implemented. Usage intentions [11,12] were measured by using 4 items for assessing participants' willingness to use the future EHR system. The job relevance [11,20] of the future EHR system to the GP's job was measured with 2 items that reflected greater perceived job relevance of the future EHR system to their work tasks. A 5-item measure was adapted from previous research [23-25] to assess the relationship between current computer and internet use for GPs' professional and personal needs and the current use of other technology with the intention to use the EHR system. Satisfaction with the current system was measured as a possible technology acceptance construct by using 5 questions for measuring participants' satisfaction with the currently used eHealth systems in the country ("My Term" and "Health Smart Card").

Questions relating to the 3 UTAUT moderators (ie, gender, age, previous work experience [8]) were also included in the questionnaire. Participants were asked to state their gender, age, and years of work experience in the current service. The voluntariness of use [8] was excluded from the questionnaire because the use of the future EHR in the country will be mandatory, so this question was redundant. The questionnaire was first developed in English and then translated into the Macedonian language using the translation back-translation method [27]. The original questionnaire and technology acceptance constructs used in the research are available from the authors on request [28].

Various approaches such as descriptive statistics, two-tailed independent sample *t* tests, Spearman rank correlations, internal consistency reliability (Cronbach α), hierarchical linear regression, moderated regression analyses, and mediation analyses were applied to analyze the collected data.

Research Ethics

Research ethics approval was obtained in accordance with the Research Ethics Policy of the University of Sheffield before commencement of the study [29]. The questionnaire was designed to avoid collecting any of the GPs' personal information. Participants were informed that they could voluntarily participate in the study.

Results

Response Rate

A total of 458 completed questionnaires were eligible for analysis, yielding a response rate of 40.2%. The age of the respondents who took part in the study ranged from 24 to 65 years (mean 44.15, SD 11.41). Two-thirds of the participants in the study (303/458, 66.2%) were females and one-third (155/458, 33.8%) were males. The work experience of the participants ranged from <1 year to 38 years of experience (mean 15.45, SD 10.40).

Reliability

The internal consistency reliability of the technology acceptance constructs used in the questionnaire was assessed using Cronbach α [19,30]. The internal consistency reliability of the measures used in the study ranged from 0.69 to 0.94, suggesting that the measures we used were reliable (Table 1).

Table 1. Spearman rank correlations.

Variable	Performance expectancy	Effort expectancy	Facilitating conditions	Job relevance	Social influence	Satisfaction	Descriptive norm	Intention
Performance expectancy								
R	N/A[a]	0.71	0.56	0.66	0.65	0.58	0.61	0.59
P value	N/A	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Effort expectancy								
R	N/A	N/A	0.68	0.69	0.66	0.61	0.57	0.68
P value	N/A	N/A	<.001	.04	<.001	<.001	<.001	<.005
Facilitating conditions								
R	N/A	N/A	N/A	0.61	0.59	0.58	0.63	0.62
P value	N/A	N/A	N/A	<.001	<.001	<.001	<.001	<.001
Job relevance								
R	N/A	N/A	N/A	N/A	0.65	0.55	0.59	0.62
P value	N/A	N/A	N/A	N/A	<.001	<.001	<.001	<.005
Social influence								
R	N/A	N/A	N/A	N/A	N/A	0.58	0.68	0.63
P value	N/A	N/A	N/A	N/A	N/A	<.04	<.001	<.001
Satisfaction								
R	N/A	N/A	N/A	N/A	N/A	N/A	0.56	0.52
P value	N/A	N/A	N/A	N/A	N/A	N/A	<.001	<.001
Descriptive norm								
R	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0.58
P value	N/A	N/A	N/A	N/A	N/A	N/A	N/A	<.001
Intention								
R	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
P value	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Mean (SD)	3.95 (1.14)	3.82 (0.87)	4.04 (0.86)	3.87 (1.04)	3.73 (1.10)	3.40 (1.09)	3.96 (1.11)	4.41 (0.91)
Cronbach α	.91	.88	.74	.69	.93	.88	.85	.94

[a]N/A: not applicable.

Bivariate Correlations

Bivariate correlations were estimated using Spearman rank-order correlation coefficients before the regression analyses. Table 1 presents the Spearman rank correlations.

The Spearman correlation showed that usage intention (the main outcome, ie, the dependent variable of this research) correlated significantly and positively with all the technology acceptance constructs (*R* coefficients=0.52-0.71) included in the study.

Descriptive Statistics

The participants in this research reported a high performance expectancy from the EHR system. They expressed a positive performance expectancy of over 50% for the system. A small minority (between 10% and 15%) was not favored. Around 18%-24% of the participants were neutral. Respondents also reported high effort expectancy from the system. They reported a positive effort expectancy of over 50% from future EHRs. A small minority (between 7% and 15%) appeared to have a negative attitude, and the neutral responses were higher (between 22% and 31%). Participants reported more than 50% positive agreement with statements on social influence constructs. A smaller minority (between 11% and 13%) reported that they did not agree with the statement, and a consistent proportion (between 30% and 33%) of participants were neutral. Participants reported that facilitating conditions are important for future use of the system. They reported over 50% positive agreement with the statements. A smaller minority of participants (between 6% and 18%) appeared not to be in favor, and 14% to 25% of respondents were neutral. Participants reported over 50% positive agreement with intention statements. A smaller minority (between 4% and 5%) appeared to have low intentions, and between 11% and 14% gave neutral scores on intentions for future use.

Gender Differences in Technology Acceptance Constructs

Independent sample *t* tests were used to assess gender differences with respect to the technology acceptance constructs. The results indicated that only significant differences were identified in performance expectancy (t_{456}=2.01; *P*=.04), wherein male GPs reported significantly higher scores (mean 4.10, SD 0.08) than their female colleagues (mean 3.87, SD 1.17).

Predicting Intentions to Use the EHR System in the Future

Hierarchical linear regression was used to assess the multivariate association between intentions to use the EHR system and UTAUT constructs. The analysis was completed in 2 steps to differentially assess the effects of demographic and information technology use/work-related constructs (entered in the first stage of the analysis) and the effects of technology acceptance constructs (the second step of the analysis). The overall model predicted (R^2) 65.4% of the variance in intention to use the future EHR system F_{445}=106.77; *P*<.001. In the first step of the analysis, only the use of other technology variables (β=−.146; *P*<.001) predicted intention to use the future EHR system. In the second step of the analysis, the addition of the UTAUT constructs significantly increased the predicted variance in intention to use the future EHR system by 63.2%. The significant predictors of intention to use the EHR system at the final step of the analysis included facilitating conditions (β=.232; *P*<.001), effort expectancy (β=.217; *P*<.001), descriptive norms (β=.198, *P*<.001), job relevance (β=.172; *P*<.001), and social influence (β=.108; *P*=.04). The results of the hierarchical regression analysis are presented in Table 2.

Table 2. Predictors of intentions to use the electronic health record system.

Steps: independent constructs	95% CI for unstandardized β weights (B)	Standard β	Adjusted R^2	*P* value
Step 1			1.8	
Age (years)	0.007 to 0.021	.091		.31
Gender	0.212 to 0.157	.014		.76
Work experience	0.022 to 0.007	.092		.29
Computer use (years)	0.016 to 0.019	.010		.38
Use of other technology	0.593 to 0.110	−.146		.004
Use of internet for personal	0.055 to 0.160	.050		.33
Use of internet for work	0.089 to 0.142	.021		.65
Step 2			65.4	
Age (years)	0.014 to 0.004	.062		.25
Gender	0.016 to 0.206	.049		.09
Work experience	0.015 to 0.003	.049		.34
Computer use (years)	0.011 to 0.011	.001		.90
Use of other technology	0.144 to 0.151	.001		.96
Use of internet for personal	0.060 to 0.069	.004		.89
Use of internet for work	0.188 to 0.048	.096		<.001
Performance expectancy	0.012 to 0.135	.076		.10
Effort expectancy	0.119 to 0.335	.217		<.001
Facilitating conditions	0.157 to 0.336	.232		<.001
Job relevance	0.070 to 0.232	.172		<.001
Social influence	0.016 to 0.162	.108		.01
Satisfaction	−0.063 to 0.064	.001		.98
Descriptive norm	0.135 to 0.282	.198		<.001

Indirect Effects of Job Relevance on Usage Intentions

We used a multiple mediation methodology [31] to assess the indirect effect of job relevance on usage intentions, after controlling for the potential mediation effects of the UTAUT constructs. Bootstrapping and bias-corrected confidence intervals were used to assess the total and indirect effects of the independent variable X (job relevance) on the dependent variable Y (usage intentions), through the effects of multiple mediators, Ms (effort expectancy; social influence, descriptive norm; and facilitating conditions). For the analysis, we used the SPSS Macro Indirect 30 with 1000 resamples and 95% CIs, and the Sobel test (*z*) was used to enable effect size comparisons between the mediators [31].

The mediation analysis showed that the association between job relevance and intentions was mediated by effort expectancy (z=5.81; P<.001), social influence (z=2.64; P=.008), descriptive norms (z=4.91; P<.001), and facilitating conditions (z=4.54; P<.001). The mediation effect of effort expectancy was significantly higher (P=.02) than the effects of social influence and descriptive norms.

Moderation Effects Between UTAUT Constructs

In total, 8 moderated regression analyses were used to assess the interactive effects of gender, age, and working experience on the relationships between the UTAUT constructs (effort expectancy, social influence, and facilitating conditions) on intentions to use the EHR system. Technology acceptance predictors were mean-centered to avoid multicollinearity [32]. As the direct effect of performance expectancy was nonsignificant, we did not assess the interaction between this variable and gender, age, and experience. An interaction term was computed (independent variable×moderator) for each pair of associations, and each moderated regression analysis was completed in 2 steps. The first step included the main effects of the independent variable and moderator, and the second step included the interaction term. Unstandardized β weights (B) and 95% CIs were estimated [32].

The analyses identified only 2 significant moderation effects. Age significantly interacted (P=.02) with social influence ($B_{age \times social\ influence}$=.005; 95% CI .001 to .010; β=.077), showing that when age was higher, the association between social influence and intentions was stronger (Figure 1). In addition, the relationship between social influence and intention to use the system was significantly moderated (P=.02) by experience ($B_{experience \times social\ influence}$=.005; 95% CI 0.001 to 0.010; β=.080), showing that among GPs in the early stages of work experience, there was a stronger relationship between the social influence and intentions to use the EHR system.

Discussion

Initial Findings

This research identified the significant correlates of technology acceptance predictors for future EHR systems among GPs in the Republic of North Macedonia. On the basis of previous research using the UTAUT in health care settings (8), it was hypothesized that UTAUT constructs (ie, performance expectancy, effort expectancy, facilitating conditions, and social influence) would be associated with intentions to use the EHR system in the future and mediate the relationship of intentions with job relevance, satisfaction with using the eHealth systems in the country, and use of other (non–health care) technology. On the basis of the UTAUT premises [8], it was further hypothesized that the associations between UTAUT constructs and usage intentions would be moderated by age, gender, and previous work experience. Finally, we anticipated that descriptive norms would provide an alternative and useful measure of social norms in the context of UTAUT and health care technologies; therefore, descriptive norms would be significantly associated with usage intentions over and above the effects of other predictors and social norms more specifically.

H1 was accepted, as effort expectancy, social influence, and facilitating conditions constructs were significantly associated with GPs' intention to use the future EHR system in the multivariate model, which accounted for 65.4% of the variance in intentions. However, although performance expectancy was significantly associated with intentions in the bivariate correlation analysis (Table 1), this association was not significant in the multivariate model. H2 was also supported, as job relevance was significantly and directly associated with usage intentions. H3 was also accepted because age and experience were reported as moderators of the social influence construct. Finally, H4 was accepted as a descriptive norm significantly associated with EHR use intentions.

These findings are in line with previous research [13,14,20], indicating a positive and significant association between effort expectancy and intentions to use health care technology among health care professionals. Although the original UTAUT model [8] posits that performance expectancy is among the strongest predictors of intention to use a system, our study did not support this contention. This is in line with previous research in the Republic of North Macedonia [22]. Facilitating conditions and job relevance were also associated with intentions in this study, and their effect as predictors on EHR intentions had only previously been reported in a limited number of studies [14].

The significant multivariate association between effort expectancy and EHR use intentions corroborates previous research on health care professionals in the Republic of North Macedonia [22]. Taken together, these findings may indicate that GPs in a specific country are not fully aware of the potential benefits of the proposed EHR system and consider perceived effort and supportive infrastructure as more relevant in their decision to use (or not use) such technology. This may explain the nonsignificant multivariate association between performance expectancy and intentions to use the future EHR system. In practical terms, this means that efforts to promote EHR use among GPs in a specific country should address the issue of the ease of using the system (ie, less effort) and the existence of supportive infrastructure, especially among GPs of older age with more years of medical practice experience.

The moderated regression analyses indicated that age and experience moderated the effects of social influence construct on intentions to use the future EHR system. However, no previous studies in these areas have used moderated regression analyses. The mediation analyses showed that the effect of job relevance was mediated by effort expectancy, social influence, descriptive norms, and facilitating conditions. This means that perceiving the use of the EHR system as relevant to GP work can only partially explain the GPs' decision to use the system. Other, more relevant considerations, such as the perceived effort in using it and the existence of relevant technical support and infrastructure, as well as the perceived use by other GPs, appear to be more prominent considerations in the decision-making process and further explain the association between job relevance and usage intentions. In other words, GPs would be willing to use health care technology that appears relevant to

their job, to the extent that this technology is seen as less effortful to use, supported by relevant infrastructure, and endorsed by more colleagues.

Principal Findings

GPs' decision to use job-relevant health care technology, such as an EHR system, is multifaceted and based on several considerations. Primarily, the perceived effortless use and the existence of supportive infrastructure appear to be highly relevant to the decision to use the EHR system in question, followed by perceptions of endorsed (and actual) use by colleagues. Taken together, these considerations appear to be more important than the perceived benefits of the EHR system in daily practice.

Implications for Design and Implementation of EHR Systems

The findings of this study may be useful for policy makers and managers when developing and implementing new ICTs in health care. The contextualized technology acceptance model developed for this study contributes to understanding the drivers of the acceptance of new technology in this country in Southeast Europe. The emphasis on the design and development of future EHRs should be easy to use. The effect of social influence on intentions to use the EHR system may be moderated by the age and experience (moderators) of the GPs.

Managers and policy makers should use workshops and tools that will persuade end users about the ease of use of EHR systems. Future EHR systems should provide effective technical support (facilitating conditions). The influence of key colleagues may facilitate the implementation of EHR systems (social influence).

Limitations

A quantitative approach was applied in this study, and it is possible that a qualitative approach may have provided a more in-depth explanation of participants' attitudes. GPs who use ICT less in their professional roles may have been underrepresented in this study, which may have created a response bias in this sample. It is also possible that the views in the research, in relation to readiness to adopt the EHR system, reflect those GPs who were more familiar with using ICT. The response rate of 40.2% is not ideal for generalizing the findings of this research to the whole GP population in the country. Although the research instrument was applied to a large sample of GPs, there was a self-selection bias among respondents. The EHR system in 2020, although planned, was not implemented in the Republic of North Macedonia. However, it is possible that the views of health care professionals, such as performance expectancy and various forms of computer and internet use, have changed over time.

Gender split and other demographic data in the GP population in the country were not available through the National Fund of Health Insurance. The gender distribution in the respondents (2/3 female vs 1/3 male) of this research cannot be compared with the GPs' gender split. Data for this research were collected in 2014, and there is a time gap with its presentation in this study.

The application of a newer technology acceptance model, such as UTAUT2, was considered a possible limitation. However, the newly added variables to the UTAUT2, such as hedonic motivation (enjoyment derived from using the technology) and price value (trade-off between perceived benefits and monetary costs) were less relevant to the planned EHR (proposed mandatory use of the EHR system reimbursed by the government).

Comparison With Prior Work

The UTAUT model has been applied in a few studies in health care settings to assess the intentions to use the EHR system. Performance expectancy and effort expectancy were found to be strong predictors, which is different from the findings of this study. The main finding of this study that only effort expectancy (not performance expectancy) was established as a predictor of intention to use the EHR system is different from those in the relevant literature [13,14]. Job relevance was assessed and proved to be a predictor of intention in this study. However, this technology acceptance construct was assessed and established as a predictor of intention among health care professionals in the relevant literature [20]. Social influence, a technology acceptance construct similar to subjective norms, has been more widely used and has been shown to be a behavioral predictor in health care settings in the relevant literature. The findings of this study, where social influence was established as a behavioral predictor, correspond with those described in the literature [14,18,20,33]. However, as performance expectancy was not established as a technology acceptance predictor of intentions in this study, there may be a gap in the awareness of its expected benefits. Therefore, the possible awareness gap may be explored in future research.

Conclusions

The modified version of the UTAUT applied in this study is a useful tool for researchers to assess attitudes and intentions to use new eHealth systems. The main findings from this study indicated that effort expectancy (not performance expectancy) and facilitating conditions (ie, perceived tech support and supportive infrastructure) were the strongest predictors of intentions for the future use of the EHR system among GPs. Taken together, the main findings of our study suggest that health care technology acceptance can be explained by models, such as the UTAUT model. However, different variables appear to predict intentions to use health care technology in different countries, suggesting that future research may address cultural and contextual influences in health care technology acceptance and that modified versions of the UTAUT may be relevant in different countries.

Conflicts of Interest

None declared.

References

1. De Pietro C, Francetic I. E-health in Switzerland: The laborious adoption of the federal law on electronic health records (EHR) and health information exchange (HIE) networks. Health Policy 2017;122(2):69-74 [FREE Full text] [doi: 10.1016/j.healthpol.2017.11.005]
2. Bahadori M, Alimohammadzadeh K, Abdolkarimi K, Ravangard R. Factors affecting physicians' attitudes towards the implementation of electronic health records using structural equation modeling (SEM). Shiraz e-Medical Journal 2017 Oct 16;18(11) [FREE Full text] [doi: 10.5812/semj.13729]
3. Short D, Frischer M, Bashford J. Barriers to the adoption of computerised decision support systems in general practice consultations: a qualitative study of GPs' perspectives. International Journal of Medical Informatics 2004;73(4):357-362 [FREE Full text] [doi: 10.1016/j.ijmedinf.2004.02.001]
4. Tubaishat A. Perceived usefulness and perceived ease of use of electronic health records among nurses: Application of technology acceptance model. Informatics for Health and Social Care 2017 Sep 18;43(4):379-389 [FREE Full text] [doi: 10.1080/17538157.2017.1363761]
5. Abdekhoda M, Ahmadi M, Gohari M, Noruzi A. The effects of organizational contextual factors on physicians' attitude toward adoption of electronic medical records. Journal of Biomedical Informatics 2015 Feb;53:174-179 [FREE Full text] [doi: 10.1016/j.jbi.2014.10.008]
6. Simon SR, Kaushal R, Cleary PD. Correlates of electronic health record adoption in office practices: A statewide survey. Journal of American Medical Informatics Association 2007 Jan;14(1):110-117 [FREE Full text] [doi: 10.1197/jamia.M2187]
7. Berg M. Implementing information systems in health care organizations: Myths and challenges. International Journal of Medical Informatics 2001 Dec;64(2-3):143-156 [FREE Full text] [doi: 10.1016/s1386-5056(01)00200-3]
8. Venkatesh V, Morris MG, Davis GB, Davis FD. User acceptance of information technology: Toward a unified view. MIS Quarterly 2003 Sep;27(3):425-478. [doi: 10.2307/30036540]
9. Hu PJ, Chau PY. Physician acceptance of telemedicine technology: an empirical investigation. Topics in Health Information Management 1999 May 01;19(4):20-35 [FREE Full text] [Medline: 10387653]
10. Hu PJ, Chau PY, Sheng ORL, Tam KY. Examining the technology acceptance model using physician acceptance of telemedicine technology. Journal of Management Information Systems 2015 Dec 02;16(2):91-112 [FREE Full text] [doi: 10.1080/07421222.1999.11518247]
11. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science 2000 Feb 01;46(2):186-204 [FREE Full text] [doi: 10.1287/mnsc.46.2.186.11926]
12. Myn YY, Jackson JD, Park JS, Probst JC. Understanding information technology acceptance by individual professionals: Toward an integrative view. Information & Management 2006 Apr;43(3):350-363 [FREE Full text] [doi: 10.1016/j.im.2005.08.006]
13. Harle CA, Devar MA. Factors in physician expectations of a forthcoming electronic health record implementation. aFactors in physician expectations of a forthcoming electronic health record implementation: proceedings of the 45th Hawaii International Conference on System Sciences. Maui, Hawaii, USA; 2012 Presented at: 45th Hawaii International International Conference on Systems Science (HICSS-45 2012); 4-7 January 2012; Koloa, Kauai, Hawaii, USA p. 2869-2878 URL: https://www.computer.org/csdl/proceedings-article/hicss/2012/4525c869/12OmNAJ4pf3 [doi: 10.1109/hicss.2012.277]
14. Razeghi RR, Nasiripour AA. An investigation of factors affecting electronic health record (EHR) in health care centres. Scholars Journal of Economics, Business and Management 2014;1(1):19-24 [FREE Full text] [doi: 10.1007/springerreference_82314]
15. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decision Sciences 2008 May 09;39(2):273-315 [FREE Full text] [doi: 10.1111/j.1540-5915.2008.00192.x]
16. Kim S, Lee K, Hwang H, Yoo S. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital. BMC Medical Informatics and Decision Making; 2016 Jan 30;16(1) [FREE Full text] [doi: 10.1186/s12911-016-0249-8]
17. Hsieh H, Kuo Y, Wang S, Chuang B, Tsai C. A study of personal health record user's behavioral model based on the PMT and UTAUT integrative perspective. International Journal of Environmental Research and Public Health 2016 Dec 23;14(1):8. [doi: 10.3390/ijerph14010008]
18. Steininger K, Stiglbauer B. EHR acceptance among Austrian resident doctors. Health Policy and Technology 2015 Jun;4(2):121-130. [doi: 10.1016/j.hlpt.2015.02.003]
19. Devine EB, Patel R, Dixon D, Sullivan S. Assessing attitudes toward electronic prescribing adoption in primary care: a survey of prescribers and staff. Journal of Innovation in Health Informatics 2010;18(3):177-187. [doi: 10.14236/jhi.v18i3.770]
20. Archer N, Cocosila M. A comparison of physician pre-adoption and adoption views on electronic health records in Canadian medical practices. Journal of Medical Internet Research 2011 Aug 12;13(3):e57 [FREE Full text] [doi: 10.2196/jmir.1726]
21. Rivis A, Sheeran P. Descriptive norms as an additional predictor in the theory of planned behaviour: A meta-analysis. Current Psychology 2003;22(3):218-233. [doi: 10.1007/s12144-003-1018-2]
22. Ketikidis P, Dimitrovski T, Bath PA, Lazuras L. Acceptance of health information technology in health professionals: an application of the revised technology acceptance model. Health Informatics Journal 2012 Jun 24;18(2):124-134 [FREE Full text] [doi: 10.1177/1460458211435425]

23. Teo T, Lee SB, Chai CS. Understanding pre‐service teachers' computer attitudes: applying and extending the technology acceptance model. Journal of Computer Assisted Learning 2007 Jul 17:128-143. [doi: 10.1111/j.1365-2729.2007.00247.x]
24. Devine EB, Patel R, Dixon D. Assessing attitudes toward electronic prescribing adoption in primary care: a survey of prescribers and staff. Journal of Innovation in Health Informatics 2010;18(3):177-187. [doi: 10.14236/jhi.v18i3.770]
25. Moton ME, Wiedenbeck S. EHR acceptance factors in ambulatory care: A survey of physician perceptions. Perspectives in Health Information Management 2010 Jan 01:1-19 [FREE Full text]
26. Holden RJ, Asan O, Wozniak EM. Nurses' perceptions, acceptance, and use of a novel in-room pediatric ICU technology: Testing an expanded technology acceptance model. BMC Medical Informatics and Decision Making 2016 Nov 15;16(1):1-10 [FREE Full text] [doi: 10.1186/s12911-016-0388-y]
27. Hambleton RK. The next generation of the ITC test translation and adaptation guidelines. European Journal of Psychological Assessment 2001 Sep;17(3):164-172. [doi: 10.1027//1015-5759.17.3.164]
28. Dimitrovski T. Investigation for national readiness for e-Health in South East European country: technology acceptance for electronic health records. White Rose e-Thesis online. 2018 Nov 21. URL: http://etheses.whiterose.ac.uk/22172/ [accessed 2019-04-13]
29. Research Ethics. University of Sheffield Ethics Policy. URL: http://www.sheffield.ac.uk/polopoly_fs/1.112642!/file/Full-Ethics-Policy.pdf [accessed 2014-01-30]
30. Fink A. The Survey Handbook. Los Angeles, USA: Sage Publications Inc; 1995:58-75.
31. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods 2008 Aug;40(3):879-891. [doi: 10.3758/brm.40.3.879]
32. Cohen P, West SG, Aiken LS. Applied multiple regression/correlation analysis for the behavioral sciences. New York: Routledge; 2014:81-105.
33. Al-Adwan SA, Berger H. Exploring physicians' behavioural intention toward the adoption of electronic health records: An empirical study from Jordan. International Journal of Healthcare Technology and Management 2016 Feb 08;15(2):89-115. [doi: 10.1504/ijhtm.2015.074538]

Abbreviations

EHR: electronic health record
GP: general practitioner
ICT: information and communication technology
UTAUT: Unified Theory of Acceptance and Use of Technology

Edited by G Eysenbach, R Kukafka; submitted 05.06.20; peer-reviewed by C Fincham, I Mircheva; comments to author 03.11.20; revised version received 29.12.20; accepted 16.01.21; published 05.04.21.

Please cite as:
Dimitrovski T, Bath PA, Ketikidis P, Lazuras L
Factors Affecting General Practitioners' Readiness to Accept and Use an Electronic Health Record System in the Republic of North Macedonia: A National Survey of General Practitioners
JMIR Med Inform 2021;9(4):e21109
URL: https://medinform.jmir.org/2021/4/e21109
doi:10.2196/21109
PMID:33818399

Original Paper

Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media Comments: Text Mining and Thematic Analysis Approach

Oladapo Oyebode[1], BSc, MSc; Chinenye Ndulue[1], BSc, MCS; Ashfaq Adib[1], BSc; Dinesh Mulchandani[1], BSc; Banuchitra Suruliraj[1], BSc, MCS; Fidelia Anulika Orji[2], BSc, MSc; Christine T Chambers[3,4], PhD, RPsych; Sandra Meier[5], PhD; Rita Orji[1], PhD

[1]Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

[2]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[3]Department of Psychology and Neuroscience, Dalhousie University, Halifax, NS, Canada

[4]Department of Pediatrics, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

[5]Department of Psychiatry, Faculty of Medicine, Dalhousie University, Halifax, NS, Canada

Corresponding Author:
Oladapo Oyebode, BSc, MSc
Faculty of Computer Science
Dalhousie University
6050 University Avenue
Halifax, NS, B3H 1W5
Canada
Phone: 1 902 494 2093
Email: oladapo.oyebode@dal.ca

Abstract

Background: The COVID-19 pandemic has caused a global health crisis that affects many aspects of human lives. In the absence of vaccines and antivirals, several behavioral change and policy initiatives such as physical distancing have been implemented to control the spread of COVID-19. Social media data can reveal public perceptions toward how governments and health agencies worldwide are handling the pandemic, and the impact of the disease on people regardless of their geographic locations in line with various factors that hinder or facilitate the efforts to control the spread of the pandemic globally.

Objective: This paper aims to investigate the impact of the COVID-19 pandemic on people worldwide using social media data.

Methods: We applied natural language processing (NLP) and thematic analysis to understand public opinions, experiences, and issues with respect to the COVID-19 pandemic using social media data. First, we collected over 47 million COVID-19–related comments from Twitter, Facebook, YouTube, and three online discussion forums. Second, we performed data preprocessing, which involved applying NLP techniques to clean and prepare the data for automated key phrase extraction. Third, we applied the NLP approach to extract meaningful key phrases from over 1 million randomly selected comments and computed sentiment score for each key phrase and assigned sentiment polarity (ie, positive, negative, or neutral) based on the score using a lexicon-based technique. Fourth, we grouped related negative and positive key phrases into categories or broad themes.

Results: A total of 34 negative themes emerged, out of which 15 were health-related issues, psychosocial issues, and social issues related to the COVID-19 pandemic from the public perspective. Some of the health-related issues were *increased mortality*, *health concerns*, *struggling health systems*, and *fitness issues*; while some of the psychosocial issues were *frustrations due to life disruptions*, *panic shopping*, and *expression of fear*. Social issues were *harassment*, *domestic violence*, and *wrong societal attitude*. In addition, 20 positive themes emerged from our results. Some of the positive themes were *public awareness*, *encouragement*, *gratitude*, *cleaner environment*, *online learning*, *charity*, *spiritual support*, and *innovative research*.

Conclusions: We uncovered various negative and positive themes representing public perceptions toward the COVID-19 pandemic and recommended interventions that can help address the health, psychosocial, and social issues based on the positive themes and other research evidence. These interventions will help governments, health professionals and agencies, institutions, and individuals in their efforts to curb the spread of COVID-19 and minimize its impact, and in reacting to any future pandemics.

(JMIR Med Inform 2021;9(4):e22734) doi:10.2196/22734

KEYWORDS

social media; COVID-19; coronavirus; infodemiology; infoveillance; natural language processing; text mining; thematic analysis; interventions; health issues; psychosocial issues; social issues

Introduction

Background

Infectious diseases have occurred in the past and continue to emerge. Infectious diseases are termed "emerging" if they newly appear in a population or have existed but are increasing rapidly in incidence or geographic range [1]. Examples of emerging infectious diseases include acquired immunodeficiency syndrome, Ebola, dengue hemorrhagic fever, Lassa fever, severe acute respiratory syndrome (SARS), H1N1 flu, Zika, etc [2]. Evidence shows that emerging infectious diseases are among the leading causes of death and disability globally [3]. For instance, a 1-year estimate of the 2009 H1N1 flu pandemic shows that 43-89 million people were infected [4], and 201,200 respiratory deaths and 83,300 cardiovascular deaths were linked to the disease [5] worldwide. In addition, 770,000 HIV deaths were recorded in 2018 alone, with approximately 37.9 million people already infected with the virus globally [6]. Ebola is another deadly infectious disease that has an average case-fatality rate of about 50%, with a range of 25%-90% case-fatality rates in past outbreaks [2,7].

In December 2019, COVID-19, caused by the novel coronavirus, emerged and soon became the latest deadly infectious disease [8,9] worldwide, with more than 9.4 million confirmed cases and over 482,800 deaths in 188 countries and regions as of June 25, 2020 [10]. Hence, it was declared a pandemic by the World Health Organization. The COVID-19 pandemic has strained the global health systems and caused socioeconomic challenges due to job losses and lockdowns (and other restrictive measures) imposed by governments and public health agencies to curtail the spread of the virus. Evidence has already shown that emerging infectious diseases impose significant burden on global economies and public health [3,11-13]. To understand public concern, personal experiences, and factors that hinder or facilitate the efforts to control the spread of the COVID-19 pandemic, social media data can produce rich and useful insights that were previously impossible in both scale and extent [14].

Over the years, social media has witnessed a surge in active users to more than 3.8 billion worldwide [15], making it a rich source of data for research in diverse domains. In the health domain, social media data (ie, user comments or posts on Twitter, Facebook, YouTube, Instagram, online forums, blogs, etc) have been used to investigate mental health issues [16,17], maternal health issues [18,19], diseases [20-24], substance use [25,26], and other health-related issues [27,28]. Other domains (eg, politics, commerce, marketing, or banking) have also witnessed widespread use of social media data to uncover new insights related to election results [29-32], election campaigns [33], customer behavior and engagement [34,35], etc. Regarding the COVID-19 crisis, social media data can reveal public perceptions toward how governments and health agencies worldwide are handling the pandemic and the social, economic, psychological, and health impacts of the disease on people regardless of their geographic locations in line with various factors that hinder or facilitate the efforts to control the spread of the COVID-19 pandemic globally.

In this paper, we apply natural language processing (NLP) to understand public opinions, experiences, and issues with respect to the COVID-19 pandemic using data from Twitter, Facebook, YouTube, and three online discussion forums (ie, *Archinect* [36,37], *LiveScience* [38], and *PushSquare* [39]). NLP is a well-established method that has been applied in many JMIR papers and other health informatics papers to understand various health-related issues. For example, Abdalla et al [40] studied the privacy implications of word embeddings trained on clinical data containing personal health information, while Bekhuis et al [41] applied NLP to extract clinical phrases and keywords from a corpus of messages posted to an internet mailing list. Specifically, we aim to answer the following research questions (RQs):

- RQ1: What are the negative issues (health, psychosocial, and social issues) shared by people on social media with respect to the COVID-19 pandemic?
- RQ2: What are the positive opinions or perceptions of people with respect to COVID-19 and how it is being handled?
- RQ3: How can the negative issues be addressed using insights from the positive opinions and other research evidence?

The methodological approach used in answering our RQs are as follows:

- We apply an NLP approach for extracting opinionated key phrases from COVID-19–related social media comments.
- We uncover various negative and positive themes, representing public perceptions toward the COVID-19 pandemic after categorizing the key phrases. Our results revealed 34 negative themes, out of which 15 were *health-related issues*, *psychosocial issues*, and *social issues* related to the pandemic from the public perspective. In addition, 20 positive themes emerged from our results.
- We recommend interventions that can help address the health, psychosocial, and social issues based on the positive themes and other research evidence. These interventions will help governments, health professionals and agencies, institutions, and individuals in their efforts to curb the spread of COVID-19 and minimize its impact, as well as in reacting to any future pandemics.

Relevant Literature

Social media has been a rich source of data for research in many domains, including health [42]. Research that uses social media in conjunction with NLP within the health domain continues to grow and cover broad application areas such as health

surveillance (eg, mental health, substance use, diseases, and pharmacovigilance), health communication, sentiment analysis, and so on [43]. For example, Park and Conway [44] used the lexicon-based approach to track prevalence of keywords indicating public interest in four health issues— Ebola, e-cigarettes, marijuana, and influenza—based on social media data. Afterward, they generated topics that explain changes in discussion volume over time using the latent Dirichlet allocation (LDA) algorithm. Similarly, Jelodar et al [45] applied LDA to extract latent topics in COVID-19–related comments and used the long short-term memory recurrent neural network technique for sentiment classification. Furthermore, Nobles et al [46] used social media data to examine the needs (including seeking health information) of the reportable sexually transmitted diseases community. Their NLP approach involves extracting the top 50 unigrams from the posts based on frequency and then generating topics using the nonnegative matrix factorization technique instead of LDA. Paul et al [47] applied the Ailment Topic Aspect Model to generate latent topics from Twitter data with the aim of detecting mentions of specific ailments including allergies, obesity, and insomnia. They used a list of key phrases to automatically identify possible systems and treatments. McNeill et al [48] investigated how the dissemination of H1N1-related advice in the United Kingdom encourages or discourages vaccine and antiviral uptake using Twitter data. They conducted an automated content analysis using the KH Coder tool (Koichi Higuchi) to explore potential topics based on frequency of occurrence and then performed a more detailed or conversational analysis to understand skepticism over economic beneficiaries of vaccination and the risks and benefits of medication based on public opinion. On the other hand, Oyebode et al [49] performed sentiment analysis on user reviews of mental health apps using the machine learning approach. They compared five classifiers (based on five different machine learning algorithms) and used the best performing classifier to predict the sentiment polarity of reviews. However, none of the aforementioned approaches considers the context in which words appear in unstructured texts, which instinctively plays a substantial role in conveying meaning.

To investigate the significance of contextual text analysis, Dave and Varma [50] compared the noncontextual n-gram chunking approach and the contextual part-of-speech (POS) chunking approach in their experimental research in the field of advertising. Although the n-gram chunking method simply extracts words of varying lengths within a sentence boundary as candidate key phrases, the POS chunking method infers the context of words using POS patterns such as one or more noun tags (NN, NNP, NNS, and NNPS) along with adjective tags (JJ) and optional cardinal tags (CD) and determiners (DT). They focused on key phrases up to a length of 6 for their experiments. Their initial assessment showed that the majority of the key phrases generated using the n-gram chunking method are not meaningful within the advertising context, hence not useful. Furthermore, they observed the impact of key phrases from both methods on the performance of classification systems based on naive Bayes, logistic regression, and bagging machine learning algorithms. Their findings revealed that systems using the POS chunking method outperformed those using the n-gram chunking method for feature extraction. We leveraged Dave and Varma's [50] contextual method in this study and extended it to capture additional POS patterns, NLP preprocessing techniques, and sentiment scoring using a lexicon-based technique.

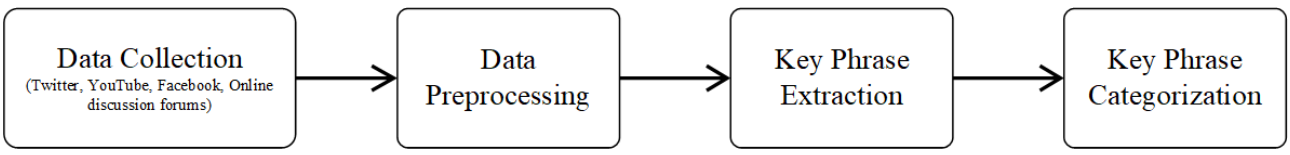
Finally, to uncover insights about the type of information shared on Twitter during the peak of the H1N1 (swine flu) pandemic in 2009, Ahmed et al [51] generated 8 broad themes using a coding method involving expert reviewers. Similarly, Bekhuis et al [41] involved two dentists to manually and iteratively classify clinical phrases into categories and subcategories. We also used this method in the key phrase categorization stage of our study to group related key phrases into categories or broad themes.

Methods

Overview

The main goal of this paper is to understand and reflect on people's personal experiences and opinions with respect to the COVID-19 pandemic using social media data. To achieve this, we applied various standard and well-known computational techniques that are highlighted in the following section and summarized in Figure 1.

Figure 1. Methodological stages.



- We collected COVID-19–related comments or posts from Twitter, Facebook, YouTube, and three online discussion forums using programming languages (Python and C#) and relevant application programming interfaces (APIs).
- We performed preprocessing tasks that involve applying NLP techniques to clean the data and prepare them for the key phrase extraction phase.
- We applied the NLP approach to extract meaningful key phrases, which are words or phrases that convey the topical content of the comments. This approach is in seven stages: *grammar definition*, *sentence breaking and tokenization*, *POS tagging*, *lemmatization*, *syntactic parsing*, *transformation and filtering*, and *sentiment scoring*.
- Based on the sentiment scores associated with the candidate key phrases, we automatically assigned sentiment polarity to the key phrases and then grouped *negative* and *positive* key phrases into categories or broad themes using the thematic analysis method. This helps to answer our RQs.

Data Collection

We used various automated techniques to collect 47,410,795 COVID-19-related or coronavirus-related comments from six

social media platforms: *Twitter*, *YouTube*, *Facebook*, *Archinect*, *LiveScience*, and *PushSquare*. The following describes the techniques and the breakdown of the data collected from each platform:

1. Twitter: We developed a tool using C# programming language to automatically extract tweets containing relevant hashtags in real time through the Twitter Streaming API [52]. To determine trending Twitter hashtags, we searched for "Trending Twitter hashtags on COVID-19" using the Google search engine and retrieved various popular hashtags from several websites including RiteTag [53] and Insider [54]. In addition, we checked a sample of top tweets on Twitter to see other common COVID-19–related hashtags they contained. The selected hashtags were *#CoronaVirus*, *#COVID-19*, *#Covid_19*, *#COVID19*, *#COVID*, *#QuarantineAndChill*, *#CoronaCrisis*, *#MyPandemicSurvivalPlan*, *#caronavirusoutbreak*, *#CoronavirusOutbreak*, *#Quarantined*, *#pandemic*, *#coronapocalypse*, *#QuarantineLife*, *#StopTheSpread*, *#CoronaVirusUpdates*, *#StayAtHome*, *#selfquarantine*, *#COVID－19*, *#panicbuying*, *#ncov2019*, *#Coronavid19*, *#SocialDistancing*, *#cronovirus*, *#CoronaVirusUpdate*, and *#CoronavirusPandemic*. A total of 47,249,973 tweets were collected between March 20 and April 3, 2020.
2. YouTube: We developed a Python script to retrieve comments associated with relevant videos through the YouTube Data API [55] using search keywords such as *covid19*, *covid-19*, and *coronavirus*. Due to YouTube's quota limits, we were only able to extract 111,722 comments across 2939 videos posted between January 1 and April 3, 2020.
3. Facebook: Due to Facebook's automated search restrictions, we applied a semiautomated approach to extract comments. First, we manually retrieved relevant groups (n=91) and pages (n=68), using the following search keywords: *COVID-19*, *Coronavirus*, and *COVID*. Afterward, we developed a Python script to extract 777 and 8382 comments posted on the groups and pages, respectively, between January 1 and April 3, 2020.
4. Online discussion forums: We developed a Python script to extract 20,747; 793; and 18,401 comments (from coronavirus-related threads) posted on *Archinect*, *LiveScience*, and *PushSquare*, respectively, between January 1 and April 3, 2020.

Data Preprocessing

Next, we applied the following NLP techniques to clean and prepare data for analysis using Python:

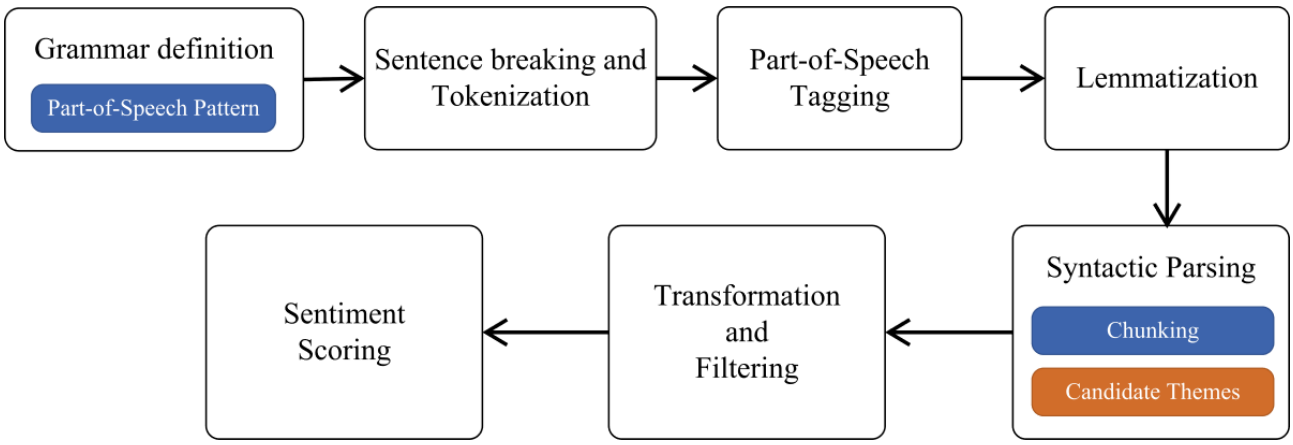
- Remove hashtags, mentions, and URLs
- Expand contractions (eg, *wouldn't* is replaced with *would not*)
- Unescape HTML characters (eg, "&" is replaced with the "&" equivalent)
- Remove HTML tags (eg, <p>, , and
)
- Remove special characters, except those with semantic implications such as periods and exclamation marks (which are useful for identifying sentence boundaries) or commas
- Reduce repeated characters (eg, *tooooooool* becomes *tool*)
- Convert slangs to their equivalent English words using online slang dictionaries [56,57], which contain 5434 entries in total
- Remove numeric words

After the preprocessing tasks were completed, non-English and duplicated comments were removed, thereby reducing the total number of comments to 8,021,341.

Key Phrase Extraction

Next, we randomly selected 1,051,616 comments (representing approximately 13% of the entire data set) and then extracted meaningful key phrases that conveyed the topical content of the comments. We refer to the data set containing the comments as *corpus* and each comment as *document* in the remaining parts of this paper. We focused on key phrases that are opinionated (ie, express or imply positive or negative sentiment [58]) since our goal was to determine public opinions and impact with respect to the COVID-19 pandemic. We extracted candidate key phrases from our corpus using a seven-stage NLP approach, shown in Figure 2. We implemented our approach using the Python programming language.

Figure 2. Natural language processing approach.



To derive meaningful key phrases, we defined the following regular grammar: *<DT>? <JJ.*>* <NN.*>* <VB.*>? (<IN>? <DT>? <JJ.*>* <NN.*>*)?* which specifies a meaningful POS pattern that the syntactic parser uses to deconstruct each sentence in the documents into their constituents [59]. Table 1 shows the various parts of speech (or syntactic categories)

captured in the grammar. These syntactic categories are based on well-established POS tagging guidelines for English [60]. In the aforementioned regular grammar, the "?" and "*" characters represent "optional" and "zero or more occurrences," respectively. Our regular grammar is aimed at generating key phrases that capture both context and sentiment of a conversation using nouns, adjectives, and verbs. Research has shown that nouns are most useful in knowing the context of a conversation (ie, what is being discussed) [61], while verbs and adjectives are important for sentiment detection [62]. Determiners and prepositions are also captured by the grammar since they usually co-occur with noun or adjective phrases (eg, a meal *for* six people or *a* hospital *on the* hilltop).

Table 1. Part-of-speech tags, description, and the corresponding matching part-of-speech pattern.

Tag	Description	Matching pattern
DT	Determiner	<DT>
JJ	Adjective	<JJ.*>
JJR	Adjective (comparative)	<JJ.*>
JJS	Adjective (superlative)	<JJ.*>
NN	Noun (singular)	<NN.*>
NNS	Noun (plural)	<NN.*>
NNP	Proper noun (singular)	<NN.*>
NNPS	Proper noun (plural)	<NN.*>
VB	Verb (base form)	<VB.*>
VBD	Verb (past tense)	<VB.*>
VBG	Verb (gerund or present participle)	<VB.*>
VBN	Verb (past participle)	<VB.*>
VBP	Verb (non–third person singular present)	<VB.*>
VBZ	Verb (third person singular present)	<VB.*>
IN	Preposition or subordinating conjunction	<IN>

Next, each document is split into sentences, and then each sentence is split into tokens or words. The sentence breaking task is achieved using an unsupervised algorithm that considers abbreviations, collocations, capitalizations, and punctuations to detect sentence boundaries [63]. The tagging module associates each token with its POS. The POS tags are based on the Penn Treebank tagset [60,64], some of which are shown in Table 1. Each token is reduced to its root form, depending on its POS. This activity is called *lemmatization*. For example, *worse* and *better*, which are both adjectives, will become *bad* and *good*, respectively. Prior to lemmatization, each token is converted to lowercase. Although Witten et al [65] applied stemming for its tokens, we chose lemmatization over stemming since lemmatization returns root words that are always meaningful and exist in the English dictionary. Stemming, on the other hand, may return root words that have no meaning at all since it merely removes prefixes or suffixes based on a rule-based method [66].

Furthermore, the syntactic parsing module deconstructs each sentence into a parse tree and then creates chunks or phrases based on the regular grammar or POS pattern defined in the first step. In other words, the parser's chunking process involves matching phrases composed of an optional determiner, zero or more of any time of adjective, zero or more of any type of noun, any type of verb (but optional), and an optional component. This component consists of an optional preposition, an optional determiner, zero or more of any type of adjective, and zero or more of any type of noun. The output of this stage is the candidate key phrases.

In the transformation and filtering stage, key phrases that are stop words (ie, words that are commonly used, such as *the*, *a*, *an*, *with*, *in*, and *that*) are removed from candidate key phrases using a predefined list $L_{stopwords}$ compiled from multiple sources (eg, [67]). We excluded negation words, which are necessary for sentiment detection, such as *not*, from the list of stop words. In addition, a subset of $L_{stopwords}$ were removed from the start and end of (and from within) the remaining key phrases in the candidate key phrases such that the meaning of the key phrases is preserved. Afterward, duplicates were removed from the candidate key phrases. Although previous research excluded key phrases above length 6 [50], we included key phrases up to length 10 in our analysis to avoid losing important key phrases that would have enriched insights from this paper. Hence, key phrases containing more than 10 words were removed from the candidate key phrases. Since our focus is on opinionated key phrases (ie, positive and negative key phrases), we applied a filtering technique that involves computing sentiment score for each key phrase and discarding nonopinionated key phrases.

Finally, to identify negative and positive key phrases in the candidate key phrases, the scoring module computes a sentiment score, S_{score}, ranging from –1 to 1 for each key phrase using the Valence Aware Dictionary for Sentiment Reasoning lexicon-based algorithm [68]. Afterward, each key phrase is

assigned a polarity (negative or positive) based on the S_{score} using the criteria recommended by Hutto and Gilbert [68]. Specifically, a key phrase is negative if S_{score}<–0.05, while a key phrase is positive if S_{score}>0.05. A neutral key phrase (with S_{score} between –0.05 and 0.05) was removed from the candidate key phrases since it is not opinionated.

Key Phrase Categorization

To answer our RQs, we categorized the final candidate key phrases into categories or broad themes using a thematic analysis approach used by Bekhuis et al [41] to classify clinical phrases into categories. In this approach, expert reviewers manually examine the key phrases and then assign them to appropriate categories. We recruited four reviewers to perform our key phrase categorization task. Specifically, we assigned the negative key phrases to a group of two reviewers (G1) and the positive key phrases to a second group of two reviewers (G2). Each reviewer independently examined the key phrases iteratively and continued to categorize related key phrases until a saturation level was reached (ie, no new categories were emerging from the key phrases). Reviewers used coding sheets in which they indicated the category each key phrase belonged to after examining it. Category names were decided by each reviewer such that a new category was created if none of the existing categories matched the key phrase being reviewed. Since key phrases are more specific than comments, the reviewers assigned each key phrase to only one category. In other words, reviewers assign a key phrase to the most appropriate category or to a new category if none of the existing categories was suitable. After categorizing the key phrases, the reviewers in each team validated each other's work and agreed or disagreed with the category assigned to each key phrase, and offered suggestions to address every disagreement. The reviewers came together after completing their validations to apply the suggestions and ensure all category names were distinct while harmonizing names that are similar. We measured interrater reliability using the *percentage agreement* metric [69]. The percentage agreement score for G1 was 98.0%, while the score for G2 was 99.3%. We refer to the categories as *themes* and the various key phrases under each category as *subthemes* in the remaining part of this paper.

Results

Key Phrase Extraction

In this section, we discuss the results of our experiments and key phrase categorization. From the large corpus used for the experiment, 427,875 negative and 520,685 positive key phrases were automatically generated. However, the majority of these key phrases were similar; hence, the reviewers reached a saturation point (during key phrase categorization) where no new categories were emerging. In total, 18,694 negative and 19,841 positive key phrases were categorized.

Negative Key Phrases

Multimedia Appendix 1 shows the top 130 negative key phrases and their dominance in terms of frequency of occurrence. Our results revealed that *death* (n=10,187) was the dominant negative key phrase, followed by *die* (n=7240), *fight* (n=5891), *bad* (n=3808), *kill* (n=3668), *lose* (n=3631), *pay* (n=3486), *leave* (n=3234), *crisis* (n=2783), *hard* (n=2720), *worry* (n=2476), *sick* (n=2314), *sad* (n=2129), and so on. More negative key phrases can be found in Multimedia Appendix 2, such as *national health emergency*, *scary time*, *life suck*, *everyone struggle*, *dangerous lie*, *child die*, *trouble breathe*, *no medicine*, *sick people*, *pay bill*, *horrible virus*, *fear coronavirus*, *extra cautious*, *steal mask*, *family die*, *people in crisis*, *bad leadership*, *in house bore*, *feel horrible*, *total incompetence*, *call virus hoax*, *conspiracy theory ridiculous*, *take no precaution*, *serious lockdown*, *increase in suicide rate*, *people starve*, *lack of preparedness*, *fight menace*, and *restriction on travel*.

Positive Key Phrases

Multimedia Appendix 3 illustrates the top 130 positive key phrases and their dominance in terms of frequency of occurrence (larger size of the gray oval represents more dominance in the figure in Multimedia Appendix 3). Our results revealed that *help* (n=18,498) was the dominant key phrase, followed by *hope* (n=7708), *protect* (n=7130), *love* (n=6895), *support* (n=6198), *good* (n=5740), *share* (n=5187), *care* (n=4917), *stay safe* (n=4917), and so on. Multimedia Appendix 4 shows more positive key phrases, such as *keep everyone safe*, *clean environment*, *trust scientific data*, *create cure*, *economic relief*, *encourage business*, *remain strong*, *good mask*, *social distancing best way*, *generous*, *respect human right*, *help prevent further spread*, *pray for health*, *social solidarity*, *support relief effort*, *protect health worker*, *good immune system*, *practice good hand hygiene*, *speak truth*, *expand testing*, *protect vulnerable people*, *free treatment*, and *ease anxiety*.

Key Phrase Categorization

Overall, 34 negative and 20 positive themes emerged after the key phrase categorization phase discussed in the Methods section. Out of the 34 negative themes, 15 were health-related, psychosocial, and social issues (which were the main focus of this paper and are shown in Tables 2-4). Table 5 shows the 15 negative themes and the corresponding number of key phrases under each theme, while Table 6 shows the negative themes and the total number of comments for each theme. Frustration due to life disruptions emerged as the top negative theme with the highest number of comments, followed by increased mortality, comparison with other diseases or incidents, nature of the disease, and harassment. On the other hand, Table 7 shows the 20 positive themes, description, and sample comments. Table 8 shows the corresponding number of key phrases under each positive theme, while Table 9 shows the total number of comments for each theme. Public awareness emerged as the top positive theme based on the number of comments, followed by spiritual support, encouragement, and charity. By identifying negative and positive themes from COVID-19–related comments, we have answered RQ1 and RQ2, respectively.

Table 2. Health-related issues: negative themes, descriptions, and corresponding sample comments.

Theme	Description	Sample comments
Increased mortality	Increasing number of deaths due to COVID-19	• "Grieving for the world and my country, every night the death count rises up. I cry for everybody that has died, for those people fighting it, & family who have lost someone. I do not know each person by name BUT I want you to know you are not alone in this pain."[a] (C58) • "...The number of deaths from the Corona-Virus in London are doubling every two days. London could end up with a worse than Italy." (C90)
Health concerns	Health concerns expressed by people, such as mental health issues (eg, anxiety, depression, stress, or obsessive-compulsive disorder), excessive drinking, migraines, fatigue, and others	• "It is been either ten or fourteen days since the nursing home I work at went on lockdown due to Covid_19 and the stress/anxiety is really starting to get to me. I am struggling to sleep at night." (C3327) • "On my fifth day of sickness, the symptoms disappeared, leaving only an odd metallic taste in my mouth, nasal mucosal ulcers and intense fatigue. This is what a former chair of the UK RCGP went through after catching COVID19." (C945)
Struggling health systems	Inability of health systems to cope with pandemic and give people adequate health care	• "What clearly shows is the correlation between countries with clean hospitals and countries with bad hospitals and corona deaths. People die with corona not of corona. In particular New York and California displays its poor health system." (C81) • "Pakistani doctors openly saying their numbers are being underreported. They claim 100s of patients in Lahore alone, and say that because of poor facilities, hospitals themselves might be spreading the virus." (C44)
Fitness issues	Inability to perform usual physical activity or attend fitness sessions and dislike for indoor workout	• "Woke up at 8:20 am and still in bed from the past 2 hours. No mood of workout" (C271) • "I just need it to be known that I hate quarantine workouts and I miss the damn gym. Also I never thought I would ever say this in life!" (C209)
Rising number of cases	Concerns over rising COVID-19 cases	• "The United States is currently on the path of the most widespread viral attack in the world." (C908) • "This has done a heinous crime on humanity by spreading to more and more areas of the country. The head of the Jamat must be booked for murder crime, as many have died due to Corona infection after attending their function." (C24)
Nature of disease	Explaining the nature of COVID-19, including its symptoms (eg, cough, loss of taste, and fever), its "no symptom" (or asymptotic) behavior, how it spreads, and vulnerable populations	• "Recent reports suggest that Covid-19 does not only affect the respiratory system, but also affects the Central Nervous System. Loss of smell and loss of taste happen to be some of the early symptoms of covid-19." (C660) • "...they are all grabbing and touching the desk. This virus is much more contagious than the flu and there is 0 immunity exposure illness. Keep doing it like this and it is only a matter of time before everyone who spoke will have the virus..." (C900) • "If 3 elders died within 72 hours and had no symptoms, should we not be testing everyone?" (C2447)
Comparison with other diseases or incidents	Comparing COVID-19 with other diseases or incidents such as flu, pneumonia, natural disasters, and war	• "A war fought with no guns or bombs, where people flee from what they do not see or is it World War 3?" (C515) • "I believe Covid19 is this mutated H5N1 avian flu virus. It is airborne, which might explain the rapid spread around the world." (C671)

[a]All comments are included verbatim, including spelling and grammatical mistakes.

Table 3. Psychosocial issues: negative themes, descriptions, and corresponding sample comments.

Theme	Description	Sample comments
Expression of fear	Expression and spread of fear among people, including fear of infection, sickness, and death	• "...Indigenous tribes are closing off their reserves to visitors as they fear the disease that is fast spreading across South America could wipe them out..."[a] (C3884) • "...This virus has caused a lot of fear in the lives of many, it has also brought about different mindset in the heart of men. Truly the world is coming to an end." (C7227)
Retrospection	People recalling past life prior to pandemic and wished they got it back	• "I miss football. I miss my family. I miss my friends. But more than all of that, I miss human touch!" (C300) • "If you consider yourself and how is social distancing going for you? I miss walking around and reading/writing at local coffee shops." (C3001)
Frustration due to life disruptions	Expression of frustrations over disruption to everyday life, such as consistent homework (for schoolers), more household chores (for moms), difficulty accessing family members or loved ones, sporting event suspension, postponement of planned trips and tours, higher food prices, and restaurants closure	• "Everyday, I wake up from a very normal dream, and realize I have to do another day in this insane world. All I want to do is go see my mom and give her a hug, but I cannot! I feel so alone. I cry every day. I cannot do this much longer." (C2905) • "Day 15 of I only leave my house for food and exercise. Living in such extremes is confusing and disorienting for my body. Every time I step outside, I become hungry and start sweating, preemptively." (C4484) • "Not getting a hair cut in February was a terrible idea. I am two hair shades away from looking like an overweight member of BTS" (C89)
Work from home complaints	Complaints about working from home during pandemic, such as distractions/disturbance, psychological stress, pain, and sleep issues	• "Is anyone else experiencing leg and knee pain from working from home too much or is it just me? If so, how have you all dealt with it?..." (C824) • "Working from home is an epic fail! I am losing it between my child, pets, monkey calls, and renovations. Wake me when this is over." (C853)
Panic shopping	People stockpiling groceries and other essential goods due to pandemic	• "This panic buying is ridiculous! Heart rending! Guess it takes a situation like to show how selfish callously indifferent we really are towards other humans / animals. Have a heart people!" (C779) • "And yet there is still no logical reason for clearing the shelves of toilet roll, depriving those who are old infirm or in poverty from accessing such necessities because the self-serving privileged have greedily taken it all away." (C4)

[a]All comments are included verbatim, including spelling and grammatical mistakes.

Table 4. Social issues: negative themes, descriptions, and corresponding sample comments.

Theme	Description	Sample comments
Wrong societal attitude	Blaming people's wrong attitude as causing or fanning the spread (eg, disobeying instructions from governments and health care professionals and eating bats and wild animals)	• "Cannot believe how irresponsible people are being in regard public health. We all have a duty of care to each other, please abide by the social distance rules"[a] (C4924) • "...We are only going to die from our own arrogance if people keep going outside gathering, believing they will never get it or infect others..." (C2557)
Domestic violence	Rise in domestic violence cases in homes	• "Of all women murdered with a gun in the US, half are killed by their intimate partners. COVID19 pandemic is causing a rise in domestic violence. Close the gun stores." (C56) • "...a police station in China received 162 reports of domestic violence in February compared to 47 for the same month in 2019. Advocates attribute this rise in cases to the lockdown." (C100)
Harassment	Harassing and blaming people from certain countries, race, or religion as responsible for the COVID-19	• "...stop being so racist against Northeastern state of India. We are not Corona. Get your damn information right. The most affected places with is not Northeast. Instead of being a racist and criticising others, why do not you be more careful?" (C1413) • "An Asian American couple in Minnesota found a racist note taped to their front door blaming them for the coronavirus" (C921)

[a]All comments are included verbatim, including spelling and grammatical mistakes.

Table 5. Negative themes and the corresponding number of subthemes.

Themes	Subthemes, n
Frustration due to life disruptions	7106
Increased mortality	4938
Comparison with other diseases/incidents	2673
Harassment	870
Panic shopping	799
Health concerns	798
Nature of disease	481
Expression of fear	338
Rising number of cases	200
Work from home complaints	109
Wrong societal attitude	105
Fitness issues	104
Domestic violence	101
Struggling health systems	48
Retrospection	24

Table 6. Negative themes and the corresponding number of comments.

Themes	Comments, n
Frustration due to life disruptions	17,535
Increased mortality	9437
Comparison with other diseases/incidents	6040
Nature of disease	1909
Harassment	1335
Health concerns	1191
Panic shopping	1047
Expression of fear	578
Rising number of cases	264
Fitness issues	151
Domestic violence	123
Work from home complaints	113
Wrong societal attitude	111
Struggling health systems	53
Retrospection	25

Table 7. Positive themes, descriptions, and corresponding sample comments.

Theme	Description	Sample comments
Gratitude	Appreciating health workers, delivery workers, farmers, pilots, security agents, other frontline workers, and the government for their active roles during the pandemic	• "Let us show our appreciation for hard work the health care professionals are doing to save lives at these thought times."[a] (C37) • "Thank you to all on the Frontline and the Key Workers, keep up the amazing work, you are doing an amazing job, keep our country going. Keep Smiling in challenging times" (C156)
Public awareness	Raising awareness of the public about general safety and control measures to limit the spread of the disease (eg, good hygiene, social distancing, staying at home, face masks, and healthy eating), addressing misinformation, providing travel guidelines, etc	• "Practicing good hygiene, like washing your hands often, with soap and water, for at least 20 seconds, is the best way to prevent the spread of Coronavirus..." (C707) • "We can help ourselves by keep practicing proper hygiene, assume anyone around us could be positive and we must keep personal distancing, wear protective gear, mask, gloves, etc. Boost our immune system by eating healthy food. Search what kind of vegetable and fruits is best to increase our body's ability to fight the virus. Sanitize groceries before bringing them inside the house..." (C16) • "If extremely necessary, stay healthy while travelling by maintaining personal hygiene, cough etiquette and keeping a distance of at least one metre from others. Here are some travel tips from World Health Organization" (C537)
Cleaner environment	Evidence of cleaner environment, including less pollution and good air quality, due to pandemic-related lockdowns	• "Coronavirus, making Earth healthy again" (C588) • "Coronavirus pandemic leading to huge drop in air pollution" (C664) • "Remember the time when we used to share the post that said, 'We have only 6 months to take action and save the environment, mend your ways or there is no way to save the earth'. Guess what? A virus saved the environment better than the most evolved beings did." (C992)
Evidence of recovery from disease	Evidence of people recovering from COVID-19 with or without treatment	• "A 95-year-old, become the oldest woman in to recover from the novel coronavirus without the need for antiviral treatment after her body showed a great reaction to the disease, doctors say." (C214) • "He was referring to the number of patients treated in the Baptist Healthcare system here. He said the numbers show 90% of their Coronavirus positive patients recover at home." (C321)
Homemade protective equipment	People's ingenuity in creating essential protective equipment (eg, face masks and face shields) themselves in their homes or community	• "When your missus tells you she put Vodka in her fairy. How was I supposed to know she is trying to make homemade hand sanitizer with the washing up liquid?" (C9335) • "82 and counting. Eager volunteers all over the city are stitching face masks to ensure that our community remains protected despite the imminent shortage of PPEs in this pandemic." (C129)
Online learning	Public engagement in online learning such as schools teaching students/pupils online and self-development by enrolling in online courses covering different domains	• "We will all be spending more time at home in the next few months due to coronavirus. I have already registered myself with number of short courses with online learning for the subjects that interest me. This could be a good time to learn something new or sharpen up your skills." (C1648) • "Despite all the news, today is the first day back to school for Florida's students virtually. Distance learning might be the new normal for a while..." (C710)
Connection with family and friends	Spending time with family, friends, and loved ones due to pandemic and lockdown	• "Yesterday my mom had a video call with her group of friends instead of going out for their regular meetup. I am proud of my mom. Who said boomers are outdated?" (C7151) • "This is not the time to be selfish. This is the time to be more present. I did a roll call this morning, calling all my friends video call, family and those I care about just to check on them..." (C2560)

Theme	Description	Sample comments
Entertainment	People accessing entertaining content online or offline, such as watching movies on streaming websites and playing games	• "Stay at Home and Stay Safe. Please share me good Amazon Prime or Netflix movies and television shows." (C93) • "Downloading the play station 1 and nintendo 64 emulators on my pc. Can get any game." (C399) • "9th grade Colin has returned and will be playing grand theft auto for 8 hours straight" (C3192)
Charity	Provision of relief packages (including donations, gifts, and fundraising) for individuals, businesses, and hospitals to ease financial burden caused by the pandemic	• "Kicking off NeighborsHelpingNeighbors here in RI with the RhodeIsland Hospitality Relief Fund - a fund aiming to help industry members directly affected by COVID19." (C3623) • "The 2 trillion emergency relief package now before the House provides the following: 1,200 checks for those earning less than 75k, plus 500 per child, unemployment benefits for 39 weeks up from 26, unemployment benefits rise by 600/week for 4 months" (C10913) • "Zlatan Ibrahimovic has created a fundraiser to gather 1m to help hospitals in Italy working to tackle the coronavirus. The striker has donated €100,000 to get the fund started." (C24)
Advocacy of increased testing	Advocacy of increased testing as a means of curbing the spread by detecting infected people and isolating them quickly	• "Studies in Iceland show that half of carriers show no symptoms. Having widespread test allows them to isolate those with the virus, and thus, the virus itself. Test, test, test." (C1333) • "US surpassing all countries in number of patients. Is it simply because they are now testing at a faster rate even on slightest suspicion? Yes, one of the ways to stop this pandemic is Testing, Testing and Testing as recommended by WHO" (C6363)
Grassroots support	Extending support to people at the local or community level during the pandemic	• "The amazing QueerCare are offering support to local mutual aid groups. They have trained volunteers and are already in contact with people who will need support over the weeks and months to come..." (C2839) • "My self, and some other good citizens residing in Lagos bought 150 pieces of pocket hand Sanitizers to distribute to people who cannot afford it or have knowledge of what is all about and in same process to tell them how to stay safe. That is our own giveaway." (C773)
Access to necessary tools	Access to tracking or communication tools/features for information dissemination or remote communication during pandemic and lockdown	• "This is the most stunning visualization of how spread around the world. A mesmerizing and terrifying display of globalization and virus spread." (C6625) • "I would like to personally thank for the private chat function during Zoom video meetings." (C2244) • "This tool is useful in identifying those who are medically more at risk of suffering complications from COVID19..." (C9233)
Spiritual support	Offering prayer of recovery for those with pandemic-related health conditions and those at risk, as well as a show of hope in challenging times	• "I pray for everyone diagnose of COVID19 swift and complete recovery in Jesus name. Dear Lord, heal the world. Take away and give everyone good health." (C9230) • "Let's pray for all African communities who live in hostels, huts, rural areas with no connectivity. No information. Innocents but will also be affected by the CoronaVirus" (C2104) • "...those who have relationship with God are less likely to become depressed than those who do not. It is because their confidence and hope is in Him so, let us trust God amid the pandemic." (C1349)
Solidarity for frontline workers	Public call for support and protection of frontline workers such as health workers and delivery workers	• "What is New York City doing to protect the workers at Amazon fulfillment center? The virus is spreading quickly among the employees..." (C3332) • "While the news of is getting more urgent by the hour, it is great to know that during a turbulent time some corporations showed their support by increasing the minute wage for frontline workers, hiring 100s of people to assist with increased demand, helping unite." (C4439)

Theme	Description	Sample comments
Development of curative solutions or treatments	Ongoing efforts by health researchers to develop vaccines or drugs to cure or treat COVID-19	• "There is a bunch of solutions being researched from I guess infusions to drugs to slow virus reproduction, to vaccines. None will be ready until a year or two. Clinical studies take time and that is how we do medical science safely." (C5759) • "Coronavirus vaccine clinical trial starts Monday, U.S. official says…But officials still say it will take a year to 18 months to fully validate any potential vaccine." (C2034)
Physical activity	Efforts made by people to stay active and fit, as well as physical activity suggestions, during lockdown or isolation	• "I know it is a bad situation but damn I am improving my fitness during this lockdown" (C6335) • "Just finished our fitness session via skype with friends. I live in Barcelona, lost track of how many days since I went outside, we have to keep the body moving though" (C9430) • "Get active every day with the kids or by yourself! GoNoodle is here to help!..." (C9364)
Encouragement	Encouraging people to stay calm as they cope with the pandemic situation, encouraging people to view the pandemic from a positive standpoint and stay productive amid the challenges, encouraging people to help others in need and not panic buy, and encouraging people to obey lockdown rules and guidelines released by governments and health professionals	• "Let us all stay calm. Give the authorities time to attend to and address public concerns..." (C3) • "Yes, we can. If we stay calm and respect the rules, together we will defeat the enemy." (C38) • "Stay calm and help each other. Be careful. Do not panic buy, and never give up!" (C164) • "We have done it before and can do it now. See the positive possibilities. Redirect the substantial energy of our frustration and turn it into positive, effective, unstoppable determination for our safe and healthy future" (C2273)
Support for remote working	People's support for the work from home measure, including adapting/coping with the challenges it brings	• "Working from home with the children of school can be challenging. We have taken a look at some of the ways you can help structure your day and stay on top of working from home with our schools closed." (C148) • "Working from home amid Coronavirus pandemic was amazing at least today. Came up with some amazing designs...We are coming up with our first property in June. Already sold out" (C53)
Innovative research	Global research efforts to create innovative products to address the pandemic, including developing interventions (eg, digital or technological interventions) that help people socially, physically, emotionally, or psychologically and to improve their overall health and wellness	• "Doctors and scientists...have designed an application to help the public monitor their symptoms and the spread of the virus in real-time with the contributing to their own vital research" (C96) • "...Well, it is a huge scientific discovery! Scientists want to use artificial intelligence technology for a quicker and cheaper COVID-19 screening..." (C197) • "The COVID19 Global Hackathon is an opportunity for developers to build software solutions that drive social impact, with the aim of tackling some of the challenges related to the current coronavirus pandemic..." (C619)
Traditional remedy	Some suggestions regarding the natural or traditional means of protecting the body from contracting the disease	• "Gargling vitamin c, vinegar, warm water, and a little bit of baking soda every 20 minutes. After 5 days, she tested negative. If you or anyone you know starts getting symptoms, this can help! Catch it early before it gets to your lungs!" (C803) • "...I am sure I have Covid_19. I believe the natural healing helped my daughter but suppressed my symptoms..." (C2777)

[a]All comments are included verbatim, including spelling and grammatical mistakes.

Table 8. Positive themes and the corresponding number of subthemes.

Themes	Subthemes, n
Public awareness	8129
Spiritual support	4139
Encouragement	4033
Entertainment	688
Gratitude	670
Charity	657
Development of curative solutions or treatments	587
Advocacy of increased testing	296
Cleaner environment	214
Evidence of recovery from disease	141
Physical activity	71
Connection with family and friends	61
Online learning	46
Access to technology tools	36
Innovative research	19
Grassroots support	17
Homemade protective equipment	14
Support for remote working	10
Solidarity for frontline workers	7
Traditional remedy	6

Table 9. Positive themes and the corresponding number of comments.

Themes	Comments, n
Public awareness	22,749
Spiritual support	12,130
Encouragement	5244
Charity	942
Entertainment	798
Gratitude	758
Development of curative solutions or treatments	653
Advocacy of increased testing	341
Physical activity	285
Cleaner environment	278
Evidence of recovery from disease	156
Connection with family and friends	70
Online learning	52
Access to technology tools	52
Innovative research	21
Grassroots support	17
Homemade protective equipment	14
Support for remote working	10
Traditional remedy	9
Solidarity for frontline workers	7

Discussion

Principal Results

In this paper, we analyzed social media comments to uncover insights regarding people's opinions and perceptions toward the COVID-19 pandemic using an NLP approach. Our empirical findings revealed negative and positive themes (see Tables 2-4 and Table 7) representing negative and positive impacts of the COVID-19 pandemic and coping mechanisms on the world population. To answer RQ3, we first discussed each of the negative issues (supported by research evidence) in this section and then suggest interventions to address the issues in a later section.

Negative Issues Surrounding the COVID-19 Pandemic

Tables 2-4 show the negative themes grouped under health-related issues, psychosocial issues, and social issues from our results. The health-related issues included *health concerns*, *increased mortality*, *struggling health systems*, *fitness issues*, *nature of disease*, *rising number of cases*, and *comparison with other diseases or incidents*. The psychosocial issues were *expression of fear*, *panic shopping*, *retrospection*, *work-from-home issues*, and *frustration due to life disruptions*. The social issues were *wrong societal attitude*, *domestic violence*, and *harassment*.

Health-Related Issues

Evidence shows a rapid increase in the number of COVID-19 cases and a high case-fatality rate of 7.2% [70]. In addition, a substantial number of patients who are infected had severe pneumonia or were critically ill [70]. Another study revealed the mental health issues experienced by people and health professionals directly impacted by the COVID-19 pandemic [71], and the global health care systems' inability to deal with the outbreak [72]. The themes under this category are discussed in the following subsections. They align with existing research and uncovered additional insights with respect to the health-related issues caused by COVID-19 and witnessed by people worldwide.

Health Concerns

Based on our findings, people experienced various mental health issues (eg, anxiety, depression, stress, or obsessive compulsive disorder) during the pandemic. This is possibly due to the length of time spent staying at home (which may be traumatic for some people while causing loneliness for others), worrying about being infected with the disease and difficult living conditions, as well as guilt on the part of health care workers who feel responsible for being unable to save their patients from death. Research confirms that worry is associated with anxiety and depression [73]. Cases of mental health disorders linked to COVID-19 have also been reported [74]. Furthermore, people expressed other concerns like excessive drinking, migraines, chest pains, mild to severe fatigue, nasal mucosal ulcers, sleep disorders, and others. The following are sample comments:

Cannot sleep. Mind is racing. Feeling anxious. [C6648]

I am so stressed and my anxiety has hit the roof. I am anxious about money and how we will cope? [C238]

This coronavirus outbreak is more stressful for the family. Doing my best to keep sanitized and safe. But, fear of the invisible killer lingers on, taking a mental toll on my mother, wife, son, who are petrified every time I walk out of the main door. [C116]

The chest pains today is beyond. It kinda have crawled up a bit and I feel like I put my hand on my heart from time to time. Very tired today. But weirdly still no fever. But I am cold and I feel sick. [C12293]

We have only been in for 3 weeks we are already feeling anxiety, depression severe, so we decided to think of some ways we can keep ourselves each other in good spirits. [C8263]

Increased Mortality

People attested to an increase in death rates in many countries across different continents including North America, Europe, Asia, the Middle East, and Australia, as shown in the following sample comments. Many countries, especially those in Africa, started reporting deaths from COVID-19 (see C1264). Our findings also revealed people of varying demographics died from the disease, including teenagers, adults, and older adults, as well as those with or without underlying health conditions (see C8837 and C940).

This is why America leads the world in the death toll already, and the pace still is not showing any signs of slackening. [C3399]

UK coronavirus death rate DOUBLES as 381 die in 24 hours and boy, 13, with no health problems becomes youngest victim [C8837]

Turkish health minister shares latest coronavirus data: 16 more people have died, bringing death toll to 108 [C9000]

Kenya has recorded the first death. According to Health Cabinet Secretary,...the 66 years old man died on 26th in the afternoon at the AghaKhan hospital intensive care unit. [C1264]

100 more UK deaths in last 24h alone. These are not all elderly co-morbid people. Among these are the young, and the fit... [C940]

Struggling Health Systems

Health systems worldwide are struggling to cope with the surge in the number of patients with COVID-19 and in most cases are unable to admit patients due to limited resources [75]. Research has shown that health care burden due to COVID-19 is associated with the increase in mortality rate [76]. As revealed in the following sample comments, our findings corroborated evidence of overstretched global health systems during this pandemic.

Hospitals turning away coronavirus patients in California. EMTALA being used to set up tents outside of hospital ER, then dumping patients without treatment or testing. I obtained recordings of standard hospital practice of rejecting Covid19 patients. [C4949]

These are not just numbers. These are people and families. These lives can be saved if the chunk of $2.2 Trillion are not used for bailing out corporations and used to fix the broken US Health System. [C3000]

Fitness Issues

Evidence argues that the prevalence of physical inactivity worldwide due to nationwide quarantines or lockdowns [77]. This was confirmed in our findings, which showed that people have trouble staying fit due to an inability to control eating habits or urges while at home and have a personal dislike for indoor-only workouts, as shown in the following comments. Physical inactivity has been linked to coronary heart disease, diabetes, stroke, and mental health issues [78-80], which, in turn, are risk factors for mortality in COVID-19 adult inpatients [81].

...severely missing my gym, missing routine, and cannot control my eating while at home. Things are getting bad. [C9002]

During this shelter in place, I was gonna eat healthy and kill some workouts. But instead I've been Guy Fieri'ing around the kitchen sampling all my quarantine food every 2 hours. At this point, which will get me first - coronavirus or a coronary? [C7182]

Nature of Disease

People expressed their opinion about the nature of COVID-19 based on their experiences and information available to them. As shown in the following sample comments, people with underlying health conditions (eg, diabetes or heart disease) are at higher risk of developing severe complications from the disease. In addition, the asymptotic attribute of COVID-19 is also discussed, and the possibility of the virus infecting some critical immune cells that may lead to the failure of sensitive organs like the lungs. People also perceived the disease as racial- or nationality–independent but seems to pose more risk to men than women. The disease is also seen as highly contagious and shows symptoms such as cough, fever, fatigue, loss of smell, muscle aches, and respiratory-related symptoms (eg, shortness of breath). These findings align with clinical evidence regarding COVID-19 [82-87].

Some WTC Health Program members with certain health conditions...may be at a higher risk of serious illness from COVID19 [C3620]

...Majority do not show symptoms while spreading Covid_19 [C10033]

...If the coronavirus infects some of the immune cells then there is a cellular catastrophe and organs like lungs are gone within hours. [C9200]

Coronavirus is a disease that pays no attention to borders, race or nationality. However, it appears COVID-19 does pose a noticeably bigger threat to men than it does to women. [C1902]

> *This is absolutely true. If you have a combination of cough, fever, problems smelling, weakness, muscle aches or shortness of breath, assume you have covid19. Don't bother getting tested. I know of many docs who don't test anymore if patient has obvious symptoms.* [C2729]

Rising Number of Cases

Our findings show that more people are getting infected with COVID-19 in many parts of the world, as shown in the following sample comments. Evidence confirms increasing numbers of COVID-19 cases in North America [88,89] and Europe [90], as well as a growing concern for vulnerable continents such as Africa [91].

> *There is rapid increase in cases of COVID19 in India...I request PM to extend the lockdown to avoid community spread.* [C1325]

> *Despite infection cases increasing at exponential rate doubling every 3 days, Trump pushes workers to risk their lives for economy...* [C6522]

Comparison With Other Diseases or Incidents

Our findings revealed that people compare COVID-19 with other diseases such as the flu (eg, Spanish flu and H1N1 swine flu) and SARS, and with more extreme incidents such as war. However, although some people tend to downplay the severity of COVID-19 (see C647), others think it is dangerous or frightening (see C922 and C45). Research has shown that COVID-19 has a higher transmissibility rate than SARS [92] and has killed more people than SARS and Middle East respiratory syndrome combined [93], thereby making it a highly contagious and lethal disease.

> *It is just another strain of flu. People with weak and have health problem it will affect different than people with stronger and not so much health problems. Media is making it sound worse...* [C647]

> *I was not at all concerned about swine flu, I was not at all concerned about Ebola, I was not at all concerned about Zika virus, but this virus I was concerned about going all the way back to January. If I had enough information to be concerned about this virus, then so did they. We are a month behind on dealing with this virus and there are no excuses, even a lay person like myself knew all the way back in January that this was bigger than anything that has come before it in my lifetime...* [C922]

> *This is a war. We need to protect ourselves and minimize unnecessary contact to avoid another Spanish flu that killed 50 million people...* [C45]

Psychosocial Issues

Expression of Fear and Panic Shopping

Based on our findings, people are fearful or scared about COVID-19, and although many expressed genuine fear (including those who had lost loved ones to the disease, contracted the disease, or had an infected family member), others attributed it to fear mongering that is further amplified by the media. As a result of this fear, many people engaged in panic buying to stockpile essential items so they can stay indoors and limit movements for some days or weeks to keep themselves and their families safe. The following are sample comments:

> *Very frightening when people who have travelled think that covid cannot affect them. Such foolish behaviour and thoughts putting all of us at risk. Those who travelled please STOP moving around and be at home.* [C8887]

> *Fear mongering through projecting number of possible deaths. The media is disgusting.* [C5559]

> *Everyone who is panic shopping is driving my family and me nuts...Everyone in our area is panic-buying groceries. We can't get noodles, rice, or really any real staple foods. We hardly have any food already. I'm kinda scared.* [C2937]

Work From Home Complaints

Furthermore, the pandemic triggered work from home (or remote work) measures to promote continuity of businesses during lockdown [94], but this may have negative implications on people's lifestyle and well-being. For example, people found consistently working from home exhausting, boring, and distracting with kids at home. In addition, people living in countries without stable electricity and strong internet found it difficult and more costly to work from home, as they have to fuel their generators and pay more for considerably good internet connectivity. Evidence has shown that people work longer hours at home than on site due to difficulty in maintaining clear delineation between work and nonwork domains [95], thereby leading to work–family conflict and strain [96]. The following are sample comments:

> *Do you find working from home exhausting? You are not alone. Why is that and how can you combat it?* [C11224]

> *This thing of working from home annoys. Waking up at 4 am to have things done before the Boy wakes up. Otherwise you will spend better part of the day watching cartoons with no work done.* [C9902]

> *It is very clear from today's program how the government is not organized or coordinated with this issue. How should one work from home with constant power cuts and bad internet?* [C8855]

Frustration Due to Life Disruptions and Retrospection

Finally, people are generally frustrated about life disruptions caused by COVID-19 (which is the top issue based on our empirical findings as shown in Table 6). Based on our findings, this frustration is mostly due to decreased leisure and interaction with friends and family, authorities' actions and inactions, and uncertainty of upcoming situations, which leads to cognitive dissonance [97], insecurity, and mental discomfort [98]. People expressed their frustrations using words reflecting anger and unhappiness or sadness, as shown in the following sample comments. Research has shown that positive emotions (eg, happiness) and life satisfaction decreased during the COVID-19 pandemic [99]. Therefore, it is unsurprising that people missed (and crave for) their prepandemic lives, in retrospection (see C377).

> *I am getting more and more angry with this current situation we are in. My favourite show has had to postpone its final episode, after already postponing series 11...* [C7776]
>
> *We are literally living in a time when arbitrary shit is more important than health, wellness and preservation of life. Entitlement. Ignorance. Selfishness. An untouchable mentality. Humanity at its absolute worst. April's going to be a painful month to live through.* [C1228]
>
> *I cannot wait until this whole thing is OVER. I miss doing nails. I miss being in my element and doing the creative things I enjoy. I have no practice hand; I have no work; I feel lost. I was just licensed in Jan!* [C377]

Social Issues

Wrong Societal Attitude

Our findings revealed disapproval and concerns about people's defiance of precautionary measures or guidelines (eg, social distancing and travel guidelines) to curb the spread of COVID-19 (see C7218 and C1444) and some people's habit of eating animals assumed to be carriers of viruses (see C4013). Research has highlighted certain factors responsible for reduced compliance with public health guidelines, such as poverty, economic dislocation, lack of compensation, and mistrust of science [100-102].

> *The only good thing about Coronavirus is that it will cull the stupid people from amongst us - those that do not take it seriously and continue to gather in public, those that go overseas to attend weddings and other events when they know the risk...* [C7218]
>
> *The public response to this crisis in the UK has been absolutely pathetic. Showing we are an entitled society who cannot handle being told what to do. Embarrassing that people cannot follow simple instruction.* [C1444]
>
> *Why is that someone who knows how dangerous eating these animals are and still go right ahead and eat. Is it that they do not have any sense or is it just irresponsible stupid people to do this then what are they crazy or just insane?* [C4013]

Domestic Violence

Furthermore, an increase in domestic violence incidents was reported as a result of the COVID-19–related lockdown, as shown in the following comments. Evidence already confirms the link between COVID-19 and the rising cases of domestic violence worldwide [103-107].

> *Cases of domestic violence in the USA has skyrocketed since the CoronaVirus forced couples to stay home together for 14 days or more.* [C9900]
>
> *While domestic violence across France increased by 32% in one week, in Paris it rose by as much as 36%.* [C13720]
>
> *What worries me more than the coronavirus, is the safety and welfare of those stuck at home with their abusers, the children witnessing domestic violence and the lonely relying on company. Staying home is not always safer.* [C7910]

Harassment

Our findings also uncovered undue harassment of people from certain cultures, races, or religious background, accusing them of spreading COVID-19. The following sample comments reveal public intimidation and racist attacks toward Chinese and Asians as well as certain Indian tribes. This aligns with evidence of widespread anti-Chinese and anti-Asian xenophobic or racist attacks, especially in the United States, both physically and on social media [108-113].

> *...This supermarket refuses to sell the food to the CHINESE! We should stop going to this supermarket! We strongly against RACISM towards Chinese people abroad! THIS MUST STOP!* [C2008]
>
> *Coronavirus has only taught me one thing; some people are so racist, especially on this platform. The amount of hate, racist comments and abuse I am seeing Chinese/Asian people get is painful.* [C5000]
>
> *Indians got racism in their blood, breath, and beyond. A Mizo girl faced racism in Pune as a woman kept covering a face whenever the Mizo girl passed by. A friend Naga from maternal side in Mumbai, got called corona virus in the middle of an empty road.* [C6660]

Recommended Interventions for Addressing the Negative Issues

In this section, we suggest interventions that can help address the negative issues while drawing insights from the positive themes and relevant research evidence. This answers our RQ3.

As lockdown and physical distancing persists, people with health concerns should be able to receive medical attention without visiting a hospital. Considering the proliferation of smartphones and the current wave of global digitization, digital interventions using mobile, artificial intelligence (AI), internet of things (IoT), and virtual reality technologies have been shown to be effective for delivering remote health care (or telehealth) to patients [114-119]. This is based on our findings under the *innovative research* positive theme (see Table 7), which revealed global research efforts to create digital interventions using emerging technologies to address the health crisis caused by COVID-19. For example, mobile apps that detect mental health issues (eg, depression and anxiety) based on phone sensors (or wearable sensors) data and self-reports using machine learning and deep learning models, and then guide users through therapeutic procedures or treatments will be useful tools during and after the pandemic. In addition, these apps should allow users to book appointments with doctors, clinicians, or therapists and access remote medical advice, diagnosis, and treatments when necessary.

In addition, data-driven surveillance systems based on AI that predict the location of the next COVID-19 outbreak can enhance the effectiveness of containment efforts, thereby slowing the spread of the disease and reducing the case-fatality rate. Furthermore, the *development of curative solutions or treatments*

(see Table 7) can be accelerated by leveraging machine learning and deep learning algorithms. For example, deep learning models can be used to predict chemical compounds that can halt viral replication and to suggest drugs that can be effective against the virus.

To address fitness issues during lockdown, *physical activity* (which is one of the positive themes in our results) programs or sessions with personalized feedback delivered through mobile apps would be helpful. Research has shown that smartphone-based health programs yield significant weight loss and increase physical activity [120]. There is also an urgent need to strengthen the global health care systems to cope with current and future pandemics through public and private investments in the health sector on an ongoing basis, such as provision of public health infrastructure that is robust and adequate for the target population and easily accessible and the provision of health insurance for everyone irrespective of financial status.

Public awareness (which emerged as the top positive theme in our findings) is also crucial for addressing negative issues arising from COVID-19 by providing timely and accurate information to people, which can be lifesaving. To reach a wider audience in an efficient manner and with less cost, public awareness can be delivered through mobile technologies, such as mobile-driven and voice-enabled conversational AI agents (or chatbots) with access to evidence-based and clinically validated resources (eg, precautionary or safety measures approved by public health agencies and organizations as well as government-approved policies or guidelines), can deliver accurate information regarding COVID-19 to people in their own native language (and in an interactive fashion) through their smartphones. These chatbots can also be made to route difficult questions to health experts for real-time feedback within the same chat session. This will help to improve people's understanding of the disease, including how it differs from other infectious diseases, and how to protect themselves and their families from getting infected with COVID-19. In addition, people will be empowered with information required to effectively respond to fear mongering, domestic violence, and harassment. Evidence already shows the deployment of multilingual chatbots for public health awareness on COVID-19 symptoms, diagnosis, and precautionary measures [121]. Furthermore, chatbots can also respond to emergencies by contacting appropriate security agencies and emergency response teams on behalf of the users. Moreover, chatbots can deliver evidence-based therapeutic interventions to people while coordinating with specialists behind the scenes where necessary.

For people with nonsmartphone devices, public health agencies can partner with telecommunication companies to deliver COVID-19–related information directly to their phones as text messages at regular intervals. Social media is another platform through which evidence-based information can be shared with the public but may be overshadowed by fake news or false information, which is mostly shared on social media [122]. Nevertheless, official COVID-19–related channels managed by (or in conjunction with) reputable international health organizations (eg, World Health Organization) or local health authorities within the social media platforms, many of which have already been deployed, provide accurate information or updates about COVID-19 cases, fatality rates, and safety measures and guidelines [123,124]. In addition, people receive location-based updates on these channels, including emergency alerts, in a timely and effective manner.

Finally, based on our findings (see Table 7), *connection with family and friends*, *encouragement*, *spiritual support*, and *charity* can help to ease people's frustrations, anxiety, and trauma (due to life disruptions caused by the pandemic) by addressing their emotional, physical, and spiritual needs. Evidence shows that psychological first aid and spiritual care can promote a sense of safety, calmness, self- and collective–efficacy, connectedness, and hope, as well as help people confront and overcome fear [125]. Therefore, people should endeavor to frequently communicate and follow up with loved ones (through direct voice or video calls or by using social media), encourage others in distress to stay calm and remain positive, identify people's immediate needs and offer necessary assistance, help people find hope and meaning, and ensure the safety and comfort of vulnerable populations.

Mobile technology can play a key role in facilitating easy access to relief packages. For instance, mobile apps can be deployed with geolocation and multilingual features to help people locate the nearest food bank and charity organizations offering assistance in their geographical area. In addition, charity organizations can effectively mobilize and deliver relief items to more people, including individuals that are indisposed, based on data collected through these apps. In addition, older adults, the sick, and those in self-isolation can indicate their condition while requesting for relief so that their items can be delivered to their doorstep instead of picking it up. These apps can further integrate with other local and international charity organizations to widen the coverage of relief efforts. Recruitment of volunteers can also take place through these apps. The use data collected can be further analyzed in real time and used to predict the communities that are in dire need of assistance using machine learning or deep learning techniques.

Limitations

In this study, we analyzed data from Twitter, Facebook, YouTube, and three discussion forums. However, people may have used other social media platforms such as Instagram and other discussion forums not covered in this study to disseminate information related to the COVID-19 pandemic. Therefore, our findings may not fully reflect the entire public's opinion on social media with respect to the pandemic. Nevertheless, to have a reasonably broad understanding of public opinions, we analyzed over 1 million social media comments compared to only a few thousand commonly analyzed in many related studies. In addition, the thematic analysis used for theme categorization may be more robust; however, the large number of key phrases rendered this process time-consuming despite filtering out many irrelevant key phrases during experimentation. Accordingly, the saturation level and subsequent review and confirmation of the theme categories from a second reviewer and coder were introduced as an acceptable compromise.

Conclusions

In this paper, we explored the impact of the COVID-19 pandemic on people worldwide using social media data. We analyzed over 1 million comments obtained from six social media platforms using a seven-stage NLP approach to extract candidate key phrases, which we further categorized into broad themes using thematic analysis. Our results revealed 34 negative themes, out of which 15 were *health-related issues*, *psychosocial issues*, and *social issues* related to the COVID-19 pandemic from the public perspective. The top health-related issues were *increased mortality*, *comparison with other diseases or incidents*, *nature of disease*, and *health concerns*, while the top psychosocial issues were *frustrations due to life disruptions*, *panic shopping*, and *expression of fear*. The top social issues were *harassment* and *domestic violence*. Besides the negative themes, 20 positive themes emerged from our results. Some of the positive themes were *public awareness*, *encouragement*, *gratitude*, *cleaner environment*, *online learning*, *charity*, *spiritual support*, and *innovative research*. We reflected on our findings and recommend interventions that can help address the health, psychosocial, and social issues based on the positive themes and other research evidence.

Digital interventions using emerging technologies such as mobile apps, AI, IoT, and virtual reality will play a major role in delivering remote health care (ie, telemedicine or telehealth) to people in the comfort of their homes, including empowering them to self-manage their health and wellness. This will help to curb the spread of COVID-19 and future infectious diseases since many people will stay away from hospitals (or clinics) to book appointments or see doctors (or other health care professionals) unless it is absolutely necessary to visit, thereby keeping health workers and patients safe. These technologies are also useful in providing timely and accurate information about COVID-19 symptoms, diagnosis, treatment, precautionary and safety measures and guidelines, and other relevant information to target audience worldwide. Finally, digital interventions and other interventions discussed in this paper can help address the emotional, physical, and spiritual needs of people who are traumatized or frustrated by the disruptions caused by the pandemic. They also inform governments, health professionals and agencies, and institutions on how to react to the current COVID-19 pandemic and future pandemics.

Acknowledgments

This research was undertaken, in part, thanks to funding from the Canada Research Chairs Program. The authors acknowledge the support of the Natural Sciences and Engineering Research Council of Canada through the Discovery Grant. They also thank the DeepSense team at Dalhousie University and Compute Canada for providing the computing infrastructure used to perform our research experiments.

Authors' Contributions

OO collected data, conducted the experiments, analyzed the results, and wrote the manuscript. CN, DM, BS, and AA collected data, categorized the themes, and reviewed the manuscript. RO is the research supervisor and reviewed the manuscript. FAO is a data analyst and researcher, and reviewed the manuscript. SM is a psychiatry and epidemiology researcher, and reviewed the manuscript. CC is a clinical psychologist and reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Top 130 negative themes and their dominance in terms of frequency of occurrence (larger size of the gray oval represents more dominance).
[PDF File (Adobe PDF File), 92 KB - medinform_v9i4e22734_app1.pdf]

Multimedia Appendix 2
Sample negative themes.
[PDF File (Adobe PDF File), 344 KB - medinform_v9i4e22734_app2.pdf]

Multimedia Appendix 3
Top 130 positive themes and their dominance in terms of frequency of occurrence (larger size of the gray oval represents more dominance).
[PDF File (Adobe PDF File), 86 KB - medinform_v9i4e22734_app3.pdf]

Multimedia Appendix 4
Sample positive themes.
[PDF File (Adobe PDF File), 321 KB - medinform_v9i4e22734_app4.pdf]

References

1. Morse SS. Factors in the emergence of infectious diseases. In: Price-Smith AT, editor. Plagues and Politics: Infectious Disease and International Policy. London: Palgrave Macmillan; 2001:8-26.
2. Kimberlin DW. Red Book: 2018-2021 Report of the Committee on Infectious Diseases. Elk Grove Village, IL: American Academy of Pediatrics; 2018:152.
3. Morens DM, Folkers GK, Fauci AS. The challenge of emerging and re-emerging infectious diseases. Nature 2004 Jul 08;430(6996):242-249 [FREE Full text] [doi: 10.1038/nature02759] [Medline: 15241422]
4. Jilani TN, Jamil RT, Siddiqui AH. H1N1 influenza. In: StatPearls. Treasure Island, FL: StatPearls Publishing; 2018.
5. Dawood FS, Iuliano AD, Reed C, Meltzer MI, Shay DK, Cheng P, et al. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. Lancet Infect Dis 2012 Sep;12(9):687-695. [doi: 10.1016/S1473-3099(12)70121-4] [Medline: 22738893]
6. HIV/AIDS. World Health Organization. 2019. URL: https://www.who.int/news-room/fact-sheets/detail/hiv-aids [accessed 2020-05-16]
7. Ebola virus disease. World Health Organization. 2020. URL: https://www.who.int/en/news-room/fact-sheets/detail/ebola-virus-disease [accessed 2020-05-16]
8. Tian H, Liu Y, Li Y, Wu C, Chen B, Kraemer MUG, et al. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. Science 2020 May 08;368(6491):638-642 [FREE Full text] [doi: 10.1126/science.abb6105] [Medline: 32234804]
9. Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature 2020 Mar;579(7798):265-269 [FREE Full text] [doi: 10.1038/s41586-020-2008-3] [Medline: 32015508]
10. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). Johns Hopkins Coronavirus Resource Center. URL: https://coronavirus.jhu.edu/map.html [accessed 2020-05-16]
11. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, et al. Global trends in emerging infectious diseases. Nature 2008 Feb 21;451(7181):990-993 [FREE Full text] [doi: 10.1038/nature06536] [Medline: 18288193]
12. Bloom DE, Black S, Rappuoli R. Emerging infectious diseases: a proactive approach. Proc Natl Acad Sci U S A 2017 Apr 18;114(16):4055-4059 [FREE Full text] [doi: 10.1073/pnas.1701410114] [Medline: 28396438]
13. Fan VY, Jamison DT, Summers LH. The inclusive cost of pandemic influenza risk. Natl Bureau Econ Res 2016. [doi: 10.3386/w22137]
14. Barbier G, Liu H. Data mining in social media. In: Aggarwal CC, editor. Social Network Data Analytics. Boston, MA: Springer; 2011:327-352.
15. Kemp S. Digital 2020: global digital overview. DataReportal. 2020. URL: https://datareportal.com/reports/digital-2020-global-digital-overview [accessed 2020-05-17]
16. Robinson P, Turk D, Jilka S, Cella M. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. Soc Psychiatry Psychiatr Epidemiol 2019 Jan;54(1):51-58 [FREE Full text] [doi: 10.1007/s00127-018-1571-5] [Medline: 30069754]
17. Guntuku SC, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH. Understanding and measuring psychological stress using social media. 2019 Presented at: Thirteenth International AAAI Conference on Web and Social Media; June 11-14, 2019; Münich, Germany URL: http://arxiv.org/abs/1811.07430
18. Golder S, Chiuve S, Weissenbacher D, Klein A, O'Connor K, Bland M, et al. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. Drug Saf 2019 Mar;42(3):389-400. [doi: 10.1007/s40264-018-0731-6] [Medline: 30284214]
19. Rezaallah B, Lewis DJ, Pierce C, Zeilhofer H, Berg B. Social media surveillance of multiple sclerosis medications used during pregnancy and breastfeeding: content analysis. J Med Internet Res 2019 Aug 07;21(8):e13003 [FREE Full text] [doi: 10.2196/13003] [Medline: 31392963]
20. Blumenthal KG, Topaz M, Zhou L, Harkness T, Sa'adon R, Bar-Bachar O, et al. Mining social media data to assess the risk of skin and soft tissue infections from allergen immunotherapy. J Allergy Clin Immunol 2019 Jul;144(1):129-134 [FREE Full text] [doi: 10.1016/j.jaci.2019.01.029] [Medline: 30721764]
21. Huang Y, Huang D, Nguyen QC. Census tract food tweets and chronic disease outcomes in the U.S., 2015 2018. Int J Environ Res Public Health 2019 Mar 18;16(6) [FREE Full text] [doi: 10.3390/ijerph16060975] [Medline: 30889911]
22. Oyebode O, Orji R. Detecting factors responsible for diabetes prevalence in Nigeria using social media and machine learning. 2019 Presented at: 15th International Conference on Network and Service Management; October 21-25, 2019; Halifax, NS. [doi: 10.23919/cnsm46954.2019.9012679]
23. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011 May 04;6(5):e19467 [FREE Full text] [doi: 10.1371/journal.pone.0019467] [Medline: 21573238]
24. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010 Nov 29;5(11):e14118 [FREE Full text] [doi: 10.1371/journal.pone.0014118] [Medline: 21124761]
25. Zhan Y, Etter J, Leischow S, Zeng D. Electronic cigarette usage patterns: a case study combining survey and social media data. J Am Med Inform Assoc 2019 Jan 01;26(1):9-18 [FREE Full text] [doi: 10.1093/jamia/ocy140] [Medline: 30544163]

26. Hassanpour S, Tomita N, DeLise T, Crosier B, Marsch LA. Identifying substance use risk based on deep neural networks and Instagram social media data. Neuropsychopharmacology 2019 Feb;44(3):487-494 [FREE Full text] [doi: 10.1038/s41386-018-0247-x] [Medline: 30356094]
27. Comito C, Forestiero A, Papuzzo G. Exploiting social media to enhance clinical decision support. In: WI '19 Companion: IEEE/WIC/ACM International Conference on Web Intelligence. 2019 Presented at: WI '19; October 2019; Thessaloniki, Greece. [doi: 10.1145/3358695.3360899]
28. Huber J, Woods T, Fushi A, Duong M, Eidelman A, Zalal A, et al. Social media research strategy to understand clinician and public perception of health care messages. JDR Clin Trans Res 2020 Jan;5(1):71-81 [FREE Full text] [doi: 10.1177/2380084419849439] [Medline: 31067411]
29. Tumasjan A, Sprenger T, Sandner P, Welpe I. Predicting elections with Twitter: what 140 characters reveal about political sentiment. 2010 Presented at: Fourth International AAAI Conference on Weblogs and Social Media; May 23-26, 2010; Washington, DC.
30. Budiharto W, Meiliana M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis. J Big Data 2018 Dec 19;5(1). [doi: 10.1186/s40537-018-0164-1]
31. Tjong E, Sang K, Bos J. Predicting the 2011 Dutch Senate election results with Twitter. 2012 Presented at: 13th Conference of the European Chapter of the Association for Computational Linguistics; April 23-27, 2012; Avignon, France p. 53-60.
32. Oyebode O, Orji R. Social media and sentiment analysis: the Nigeria Presidential Election 2019. 2019 Presented at: 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference; October 17-19, 2019; Vancouver, BC p. 140-146. [doi: 10.1109/iemcon.2019.8936139]
33. Bright J, Hale S, Ganesh B, Bulovsky A, Margetts H, Howard P. Does Campaigning on Social Media Make a Difference? Evidence From Candidate Use of Twitter During the 2015 and 2017 U.K. Elections. Commun Res 2019 Sep 11;47(7):988-1009. [doi: 10.1177/0093650219872394]
34. Ma J, Tse YK, Wang X, Zhang M. Examining customer perception and behaviour through social media research – an empirical study of the United Airlines overbooking crisis. Transportation Res Part E Logistics Transportation Rev 2019 Jul;127:192-205. [doi: 10.1016/j.tre.2019.05.004]
35. Ibrahim N, Wang X. Decoding the sentiment dynamics of online retailing customers: time series analysis of social media. Comput Hum Behav 2019 Jul;96:32-45. [doi: 10.1016/j.chb.2019.02.004]
36. corona virus covid-19 and you. Archinect. URL: https://archinect.com/forum/thread/150187455/corona-virus-covid-19-and-you [accessed 2020-04-02]
37. COVID - 19 Thread Central. Archinect. URL: https://archinect.com/forum/thread/150188615/covid-19-thread-central [accessed 2020-04-02]
38. Live Science Forums. URL: https://forums.livescience.com/forums/coronavirus-epidemiology.42/ [accessed 2020-04-02]
39. Topic: corona virus panic/discussion thread. Push Square. URL: https://www.pushsquare.com/forums/ps_general_discussion/corona_virus_panicdiscussion_thread [accessed 2020-04-02]
40. Abdalla M, Abdalla M, Hirst G, Rudzicz F. Exploring the privacy-preserving properties of word embeddings: algorithmic validation study. J Med Internet Res 2020 Jul 15;22(7):e18055 [FREE Full text] [doi: 10.2196/18055] [Medline: 32673230]
41. Bekhuis T, Kreinacke M, Spallek H, Song M, O'Donnell JA. Using natural language processing to enable in-depth analysis of clinical messages posted to an internet mailing list: a feasibility study. J Med Internet Res 2011 Nov 23;13(4):e98 [FREE Full text] [doi: 10.2196/jmir.1799] [Medline: 22112583]
42. Grajales F, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
43. Conway M, Hu M, Chapman WW. Recent advances in using natural language processing to address public health research questions using social media and consumer generated data. Yearb Med Inform 2019 Aug;28(1):208-217 [FREE Full text] [doi: 10.1055/s-0039-1677918] [Medline: 31419834]
44. Park A, Conway M. Tracking health related discussions on Reddit for public health applications. AMIA Annu Symp Proc 2017;2017:1362-1371 [FREE Full text] [Medline: 29854205]
45. Jelodar H, Wang Y, Orji R, Huang S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. IEEE J Biomed Health Inform 2020 Oct;24(10):2733-2742. [doi: 10.1109/JBHI.2020.3001216] [Medline: 32750931]
46. Nobles A, Dreisbach C, Keim-Malpass J, Barnes L. "Is this a STD? Please help!": online information seeking for sexually transmitted diseases on reddit. Proc Int AAAI Conf Weblogs Soc Media 2018 Jun;2018:660-663 [FREE Full text] [Medline: 30984474]
47. Paul M, Dredze M. You are what you tweet: analyzing Twitter for public health. 2011 Presented at: Fifth International AAAI Conference on Weblogs and Social Media; July 17-21, 2011; Barcelona, Catalonia, Spain.
48. McNeill A, Harris PR, Briggs P. Twitter influence on UK vaccination and antiviral uptake during the 2009 H1N1 pandemic. Front Public Health 2016;4:26. [doi: 10.3389/fpubh.2016.00026] [Medline: 26942174]
49. Oyebode O, Alqahtani F, Orji R. Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews. IEEE Access 2020;8:111141-111158. [doi: 10.1109/access.2020.3002176]

50. Dave KS, Varma V. Pattern based keyword extraction for contextual advertising. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. 2010 Presented at: CIKM '10; October 2010; Toronto, ON p. 1885-1888.
51. Ahmed W, Bath PA, Sbaffi L, Demartini G. Novel insights into views towards H1N1 during the 2009 pandemic: a thematic analysis of Twitter data. Health Info Libr J 2019 Mar;36(1):60-72. [doi: 10.1111/hir.12247] [Medline: 30663232]
52. Consuming streaming data. Twitter Developer. URL: https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data [accessed 2020-05-20]
53. Popular hashtags for coronavirus on Twitter and Instagram. RiteTag. URL: https://ritetag.com/best-hashtags-for/coronavirus [accessed 2020-03-18]
54. Lustig H. Increasingly dark internet trends reveal a growing cynicism about the coronavirus pandemic and our future. Insider. 2020. URL: https://www.insider.com/coronavirus-twitter-hashtags-reveal-growing-cynicism-about-pandemic-survival-2020-3 [accessed 2020-03-18]
55. YouTube Data API. Google Developers. URL: https://developers.google.com/youtube/v3 [accessed 2020-05-20]
56. Slang words dictionary. GitHub. URL: https://raw.githubusercontent.com/sifei/Dictionary-for-Sentiment-Analysis/master/slang/acrynom.csv [accessed 2019-06-19]
57. Slang lookup table. GitHub. URL: https://raw.githubusercontent.com/felipebravom/StaticTwitterSent/master/extra/SentiStrength/SlangLookupTable.txt [accessed 2019-06-19]
58. Liu B. Sentence subjectivity and sentiment classification. Sentiment Analysis Mining Opinions Sentiments Emotions 2015:70-89. [doi: 10.1017/cbo9781139084789.005]
59. Nelson G. English: An Essential Grammar. England: Taylor & Francis; 2019.
60. Santorini B. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). Tech Rep 1990:570. [doi: 10.1093/oxfordhb/9780199276349.013.0011]
61. Asmuth JA, Gentner D. Context sensitivity of relational nouns. Proc Annu Meeting Cognitive Sci Soc 2005;27(27):163-168.
62. Chesley P, Vincent B, Xu L, Srihari RK. Using verbs and adjectives to automatically classify blog sentiment. Training 2006;580(263):233.
63. nltk.tokenize package. NLTK 3.5 documentation. URL: http://www.nltk.org/api/nltk.tokenize.html?highlight=tokenizer#nltk.tokenize.punkt.PunktSentenceTokenizer [accessed 2020-05-23]
64. Taylor A, Marcus M, Santorini B. The Penn Treebank: an overview. In: Treebanks: Building and Using Parsed Corpora. Dordrecht: Springer; 2003:5-22.
65. Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: practical automatic keyphrase extraction. arXiv. 1999 Feb 04. URL: http://arxiv.org/abs/cs/9902007 [accessed 2020-06-24]
66. Han P, Shen S, Wang D, Liu Y. The influence of word normalization in English document clustering. 2012 Presented at: 2012 IEEE International Conference on Computer Science and Automation Engineering; May 25-27, 2012; Zhangjiajie, China p. 116-120. [doi: 10.1109/csae.2012.6272740]
67. sebleier/NLTK's list of english stopwords. GitHub Gists. URL: https://gist.github.com/sebleier/554280 [accessed 2020-05-26]
68. Hutto C, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. 2014 Presented at: Eighth International AAAI Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, MI URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109
69. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
70. Onder G, Rezza G, Brusaferro S. Case-fatality rate and characteristics of patients dying in relation to COVID-19 in Italy. JAMA 2020 May 12;323(18):1775-1776. [doi: 10.1001/jama.2020.4683] [Medline: 32203977]
71. Xiang Y, Yang Y, Li W, Zhang L, Zhang Q, Cheung T, et al. Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed. Lancet Psychiatry 2020 Mar;7(3):228-229 [FREE Full text] [doi: 10.1016/S2215-0366(20)30046-8] [Medline: 32032543]
72. Hick JL, Biddinger PD. Novel coronavirus and old lessons - preparing the health system for the pandemic. N Engl J Med 2020 May 14;382(20):e55. [doi: 10.1056/NEJMp2005118] [Medline: 32212515]
73. Dar KA, Iqbal N, Mushtaq A. Intolerance of uncertainty, depression, and anxiety: examining the indirect and moderating effects of worry. Asian J Psychiatr 2017 Oct;29:129-133. [doi: 10.1016/j.ajp.2017.04.017] [Medline: 29061409]
74. Yao H, Chen J, Xu Y. Patients with mental health disorders in the COVID-19 epidemic. Lancet Psychiatry 2020 Apr;7(4):e21 [FREE Full text] [doi: 10.1016/S2215-0366(20)30090-0] [Medline: 32199510]
75. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. Fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med 2020 May 21;382(21):2049-2055. [doi: 10.1056/NEJMsb2005114] [Medline: 32202722]
76. Ji Y, Ma Z, Peppelenbosch MP, Pan Q. Potential association between COVID-19 mortality and health-care resource availability. Lancet Glob Health 2020 Apr;8(4):e480 [FREE Full text] [doi: 10.1016/S2214-109X(20)30068-1] [Medline: 32109372]
77. Hall G, Laddu DR, Phillips SA, Lavie CJ, Arena R. A tale of two pandemics: how will COVID-19 and global trends in physical inactivity and sedentary behavior affect one another? Prog Cardiovasc Dis 2021;64:108-110 [FREE Full text] [doi: 10.1016/j.pcad.2020.04.005] [Medline: 32277997]

78. Kivimäki M, Singh-Manoux A, Pentti J, Sabia S, Nyberg ST, Alfredsson L, IPD-Work consortium. Physical inactivity, cardiometabolic disease, and risk of dementia: an individual-participant meta-analysis. BMJ 2019 Apr 17;365:l1495 [FREE Full text] [doi: 10.1136/bmj.l1495] [Medline: 30995986]
79. Brooks SK, Webster RK, Smith LE, Woodland L, Wessely S, Greenberg N, et al. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Lancet 2020 Mar 14;395(10227):912-920 [FREE Full text] [doi: 10.1016/S0140-6736(20)30460-8] [Medline: 32112714]
80. Lippi G, Henry BM, Sanchis-Gomar F. Physical inactivity and cardiovascular disease at the time of coronavirus disease 2019 (COVID-19). Eur J Prev Cardiol 2020 Jun;27(9):906-908 [FREE Full text] [doi: 10.1177/2047487320916823] [Medline: 32270698]
81. Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet 2020 Mar 28;395(10229):1054-1062 [FREE Full text] [doi: 10.1016/S0140-6736(20)30566-3] [Medline: 32171076]
82. He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, et al. Temporal dynamics in viral shedding and transmissibility of COVID-19. Nat Med 2020 May;26(5):672-675. [doi: 10.1038/s41591-020-0869-5] [Medline: 32296168]
83. Jin J, Bai P, He W, Wu F, Liu X, Han D, et al. Gender differences in patients with COVID-19: focus on severity and mortality. Front Public Health 2020;8:152. [doi: 10.3389/fpubh.2020.00152] [Medline: 32411652]
84. Bai Y, Yao L, Wei T, Tian F, Jin D, Chen L, et al. Presumed asymptomatic carrier transmission of COVID-19. JAMA 2020 Apr 14;323(14):1406-1407 [FREE Full text] [doi: 10.1001/jama.2020.2565] [Medline: 32083643]
85. Cao X. COVID-19: immunopathology and its implications for therapy. Nat Rev Immunol 2020 May;20(5):269-270 [FREE Full text] [doi: 10.1038/s41577-020-0308-3] [Medline: 32273594]
86. Zheng Y, Ma Y, Zhang J, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol 2020 May;17(5):259-260 [FREE Full text] [doi: 10.1038/s41569-020-0360-5] [Medline: 32139904]
87. CDC COVID-19 Response Team. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019 - United States, February 12-March 28, 2020. MMWR Morb Mortal Wkly Rep 2020 Apr 03;69(13):382-386. [doi: 10.15585/mmwr.mm6913e2] [Medline: 32240123]
88. Scarabel F, Pellis L, Bragazzi NL, Wu J. Canada needs to rapidly escalate public health interventions for its COVID-19 mitigation strategies. Infect Dis Model 2020;5:316-322 [FREE Full text] [doi: 10.1016/j.idm.2020.03.004] [Medline: 32518882]
89. CDC COVID-19 Response Team. Geographic differences in COVID-19 cases, deaths, and incidence - United States, February 12-April 7, 2020. MMWR Morb Mortal Wkly Rep 2020 Apr 17;69(15):465-471. [doi: 10.15585/mmwr.mm6915e4] [Medline: 32298250]
90. Yuan J, Li M, Lv G, Lu ZK. Monitoring transmissibility and mortality of COVID-19 in Europe. Int J Infect Dis 2020 Jun;95:311-315 [FREE Full text] [doi: 10.1016/j.ijid.2020.03.050] [Medline: 32234343]
91. Kobia F, Gitaka J. COVID-19: are Africa's diagnostic challenges blunting response effectiveness? AAS Open Res 2020;3:4 [FREE Full text] [doi: 10.12688/aasopenres.13061.1] [Medline: 32399515]
92. Liu Y, Gayle AA, Wilder-Smith A, Rocklöv J. The reproductive number of COVID-19 is higher compared to SARS coronavirus. J Travel Med 2020 Mar 13;27(2) [FREE Full text] [doi: 10.1093/jtm/taaa021] [Medline: 32052846]
93. Mahase E. Coronavirus covid-19 has killed more people than SARS and MERS combined, despite lower case fatality rate. BMJ 2020 Feb 18;368:m641. [doi: 10.1136/bmj.m641] [Medline: 32071063]
94. Gottlieb C, Grobovsek J, Poschke M. Working from home across countries. Covid Economics 2020;1(8):71-91.
95. Dockery M, Bawa S. Working from home in the COVID-19 lockdown. Bankwest Curtin Economics Centre. 2020 May. URL: https://bcec.edu.au/assets/2020/05/BCEC-COVID19-Brief-4_Working-from-home.pdf [accessed 2020-06-22]
96. Thomas LT, Ganster DC. Impact of family-supportive work variables on work-family conflict and strain: a control perspective. J Appl Psychol 1995;80(1):6-15. [doi: 10.1037/0021-9010.80.1.6]
97. Foulds GA. A theory of cognitive dissonance. Br J Psychiatry 1963;109(458):164-165.
98. Tsai J, Ford ES, Li C, Zhao G, Balluz LS. Physical activity and optimal self-rated health of adults with and without diabetes. BMC Public Health 2010 Jun 23;10:365 [FREE Full text] [doi: 10.1186/1471-2458-10-365] [Medline: 20573237]
99. Li S, Wang Y, Xue J, Zhao N, Zhu T. The impact of COVID-19 epidemic declaration on psychological consequences: a study on active Weibo users. Int J Environ Res Public Health 2020 Mar 19;17(6) [FREE Full text] [doi: 10.3390/ijerph17062032] [Medline: 32204411]
100. Bodas M, Peleg K. Self-isolation compliance in the COVID-19 era influenced by compensation: findings from a recent survey in Israel. Health Aff (Millwood) 2020 Jun;39(6):936-941. [doi: 10.1377/hlthaff.2020.00382] [Medline: 32271627]
101. Wright AL, Sonin K, Driscoll J, Wilson J. Poverty and economic dislocation reduce compliance with COVID-19 shelter-in-place protocols. J Econ Behav Organ 2020 Dec;180:544-554. [doi: 10.1016/j.jebo.2020.10.008] [Medline: 33100443]
102. Plohl N, Musil B. Modeling compliance with COVID-19 prevention guidelines: the critical role of trust in science. Psychol Health Med 2021 Jan;26(1):1-12. [doi: 10.1080/13548506.2020.1772988] [Medline: 32479113]
103. Bradbury-Jones C, Isham L. The pandemic paradox: the consequences of COVID-19 on domestic violence. J Clin Nurs 2020 Jul;29(13-14):2047-2049 [FREE Full text] [doi: 10.1111/jocn.15296] [Medline: 32281158]

104. Taub A. A new Covid-19 crisis: domestic abuse rises worldwide. Chester County Community Foundation. 2020. URL: https://chescocf.org/wp-content/uploads/2020/04/Domestic-Abuse-Rises-Worldwide-New-York-Times.pdf [accessed 2020-06-21]
105. Mazza M, Marano G, Lai C, Janiri L, Sani G. Danger in danger: interpersonal violence during COVID-19 quarantine. Psychiatry Res 2020 Jul;289:113046 [FREE Full text] [doi: 10.1016/j.psychres.2020.113046] [Medline: 32387794]
106. Piquero AR, Riddell JR, Bishopp SA, Narvey C, Reid JA, Piquero NL. Staying home, staying safe? A short-term analysis of COVID-19 on Dallas domestic violence. Am J Crim Justice 2020 Jun 14:1-35 [FREE Full text] [doi: 10.1007/s12103-020-09531-7] [Medline: 32837161]
107. Boserup B, McKenney M, Elkbuli A. Alarming trends in US domestic violence during the COVID-19 pandemic. Am J Emerg Med 2020 Dec;38(12):2753-2755 [FREE Full text] [doi: 10.1016/j.ajem.2020.04.077] [Medline: 32402499]
108. Tessler H, Choi M, Kao G. The anxiety of being Asian American: hate crimes and negative biases during the COVID-19 pandemic. Am J Crim Justice 2020 Jun 10:1-11 [FREE Full text] [doi: 10.1007/s12103-020-09541-5] [Medline: 32837158]
109. Crockett D, Grier SA. Race in the marketplace and COVID-19. J Public Policy Marketing 2020 May 28;40(1):89-91. [doi: 10.1177/0743915620931448]
110. Hu J, Wang M, Lu F. COVID-19 and Asian American Pacific Islanders. J Gen Intern Med 2020 Sep;35(9):2763-2764 [FREE Full text] [doi: 10.1007/s11606-020-05953-5] [Medline: 32533432]
111. Chen HA, Trinh J, Yang GP. Anti-Asian sentiment in the United States - COVID-19 and history. Am J Surg 2020 Sep;220(3):556-557 [FREE Full text] [doi: 10.1016/j.amjsurg.2020.05.020] [Medline: 32425201]
112. Torkamaan H, Ziegler J. Rating-based preference elicitation for recommendation of stress intervention. In: Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. 2019 Presented at: UMAP '19; June 2019; Larnaca, Cyprus p. 46-50. [doi: 10.1145/3320435.3324990]
113. Ziems C, He B, Soni S, Kumar S. Racism is a virus: anti-Asian hate and counterhate in social media during the COVID-19 crisis. arXiv. Preprint posted online May 25, 2020. [FREE Full text]
114. Singh RP, Javaid M, Kataria R, Tyagi M, Haleem A, Suman R. Significant applications of virtual reality for COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):661-664 [FREE Full text] [doi: 10.1016/j.dsx.2020.05.011] [Medline: 32438329]
115. Ha SW, Kim J. Designing a scalable, accessible, and effective mobile app based solution for common mental health problems. Int J Human–Computer Interaction 2020 Apr 26;36(14):1354-1367. [doi: 10.1080/10447318.2020.1750792]
116. Torous J, Keshavan M. COVID-19, mobile health and serious mental illness. Schizophr Res 2020 Apr;218:36-37 [FREE Full text] [doi: 10.1016/j.schres.2020.04.013] [Medline: 32327314]
117. Ellahham S. Artificial intelligence: the future for diabetes care. Am J Med 2020 Aug;133(8):895-900. [doi: 10.1016/j.amjmed.2020.03.033] [Medline: 32325045]
118. Bachtiger P, Plymen CM, Pabari PA, Howard JP, Whinnett ZI, Opoku F, et al. Artificial intelligence, data sensors and interconnectivity: future opportunities for heart failure. Card Fail Rev 2020 Mar;6:e11 [FREE Full text] [doi: 10.15420/cfr.2019.14] [Medline: 32514380]
119. Hoffman DA. Increasing access to care: telehealth during COVID-19. J Law Biosci 2020;7(1):lsaa043 [FREE Full text] [doi: 10.1093/jlb/lsaa043] [Medline: 32843985]
120. Kim H, Seo K. Smartphone-based health program for improving physical activity and tackling obesity for young adults: a systematic review and meta-analysis. Int J Environ Res Public Health 2019 Dec 18;17(1) [FREE Full text] [doi: 10.3390/ijerph17010015] [Medline: 31861359]
121. Ali MY, Bhatti R. COVID-19 (coronavirus) pandemic: information sources channels for the public health awareness. Asia Pac J Public Health 2020 May;32(4):168-169 [FREE Full text] [doi: 10.1177/1010539520927261] [Medline: 32429681]
122. Ashrafi-Rizi H, Kazempour Z. Information typology in coronavirus (COVID-19) crisis; a commentary. Arch Acad Emerg Med 2020;8(1):e19 [FREE Full text] [Medline: 32185370]
123. Coronavirus (COVID-19) Information Center. Facebook. URL: https://www.facebook.com/coronavirus_info [accessed 2020-07-13]
124. COVID-19 information and resources. Google. URL: https://www.google.com/covid19/ [accessed 2020-07-13]
125. Clark PY, Joseph DM, Humphreys J. Cultural, psychological, and spiritual dimensions of palliative care in humanitarian crises. In: Waldman E, Glass M, editors. A Field Manual for Palliative Care in Humanitarian Crises. Oxford, England: Oxford University Press; 2019:1-160.

Abbreviations

AI: artificial intelligence
API: application programming interface
IoT: internet of things
LDA: latent Dirichlet allocation
NLP: natural language processing
POS: part of speech

RQ: research question
SARS: severe acute respiratory syndrome

Edited by C Lovis; submitted 21.07.20; peer-reviewed by SK Mukhiya, A Sesagiri Raamkumar, X Huang; comments to author 01.09.20; revised version received 22.10.20; accepted 25.02.21; published 06.04.21.

Please cite as:
Oyebode O, Ndulue C, Adib A, Mulchandani D, Suruliraj B, Orji FA, Chambers CT, Meier S, Orji R
Health, Psychosocial, and Social Issues Emanating From the COVID-19 Pandemic Based on Social Media Comments: Text Mining and Thematic Analysis Approach
JMIR Med Inform 2021;9(4):e22734
URL: https://medinform.jmir.org/2021/4/e22734
doi:10.2196/22734
PMID:33684052

Original Paper

A Natural Language Processing–Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study

Raffaello Furlan[1,2*], MD; Mauro Gatti[3*], PhD; Roberto Menè[1,3], MD; Dana Shiffer[1], MD; Chiara Marchiori[4], PhD; Alessandro Giaj Levra[1]; Vincenzo Saturnino[3], MS; Enrico Brunetta[1,2], MD, PhD; Franca Dipaola[1,2], MD

[1]Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

[2]Internal Medicine, Humanitas Clinical and Research Center, IRCCS, Rozzano, Milan, Italy

[3]Active Intelligence Center, IBM, Bologna, Italy

[4]IBM Research, Zurich, Switzerland

[*]these authors contributed equally

Corresponding Author:
Raffaello Furlan, MD
Department of Biomedical Sciences
Humanitas University
Via R Levi Montalcini, 4
Pieve Emanuele, Milan, 20090
Italy
Phone: 39 0282247228
Email: raffaello.furlan@hunimed.eu

Abstract

Background: Shortage of human resources, increasing educational costs, and the need to keep social distances in response to the COVID-19 worldwide outbreak have prompted the necessity of clinical training methods designed for distance learning. Virtual patient simulators (VPSs) may partially meet these needs. Natural language processing (NLP) and intelligent tutoring systems (ITSs) may further enhance the educational impact of these simulators.

Objective: The goal of this study was to develop a VPS for clinical diagnostic reasoning that integrates interaction in natural language and an ITS. We also aimed to provide preliminary results of a short-term learning test administered on undergraduate students after use of the simulator.

Methods: We trained a Siamese long short-term memory network for anamnesis and NLP algorithms combined with Systematized Nomenclature of Medicine (SNOMED) ontology for diagnostic hypothesis generation. The ITS was structured on the concepts of knowledge, assessment, and learner models. To assess short-term learning changes, 15 undergraduate medical students underwent two identical tests, composed of multiple-choice questions, before and after performing a simulation by the virtual simulator. The test was made up of 22 questions; 11 of these were core questions that were specifically designed to evaluate clinical knowledge related to the simulated case.

Results: We developed a VPS called Hepius that allows students to gather clinical information from the patient's medical history, physical exam, and investigations and allows them to formulate a differential diagnosis by using natural language. Hepius is also an ITS that provides real-time step-by-step feedback to the student and suggests specific topics the student has to review to fill in potential knowledge gaps. Results from the short-term learning test showed an increase in both mean test score (P<.001) and mean score for core questions (P<.001) when comparing presimulation and postsimulation performance.

Conclusions: By combining ITS and NLP technologies, Hepius may provide medical undergraduate students with a learning tool for training them in diagnostic reasoning. This may be particularly useful in a setting where students have restricted access to clinical wards, as is happening during the COVID-19 pandemic in many countries worldwide.

(JMIR Med Inform 2021;9(4):e24073) doi:10.2196/24073

KEYWORDS

COVID-19; intelligent tutoring system; virtual patient simulator; natural language processing; artificial intelligence; clinical diagnostic reasoning

Introduction

Learning clinical diagnostic reasoning is a critical challenge for medical students, as fallacies in diagnostic reasoning may lead to patient mistreatment with negative consequences on patient health and health care costs [1]. Adequate training and coaching are pivotal aspects for the proper development of diagnostic skills. In medical schools, clinical coaching is currently performed under the direct supervision of senior doctors, mostly in the wards [2].

Constraints in human resources and increases in educational costs prompted the development of sustainable systems for optimizing medical student tutoring [3]. In addition, the strict need to keep social distances due to the recent COVID-19 worldwide outbreak has resulted in the temporary closure of universities in many countries and denied medical students from accessing clinical wards [4,5]. From an educational standpoint, this promotes the need for clinical training methods that do not require bedside didactic activities and that do not necessarily entail continuous direct supervision by experienced doctors [6,7]. Examples of these methods are simulators, which were developed not only to support learning of specific medical procedures, such as laparoscopy [8], but also to train students in clinical diagnostic reasoning as with virtual patient simulators (VPSs) [9]. A VPS is a computer program that simulates real-life clinical scenarios, enabling students to emulate the role of a doctor by obtaining a medical history, performing a physical exam, and making diagnostic and therapeutic decisions [10]. These computer-based simulators may complement traditional training techniques without requiring direct ward attendance [11].

Previous studies based on intelligent tutoring systems (ITSs) [12] have shown the effectiveness of programs [13] specifically developed to teach and practice knowledge in several areas, including mathematics and physics [14]. ITS technologies can be adapted to students' specific learning needs, thus potentially increasing the simulator's teaching effectiveness [15-17]. Natural language processing (NLP) may complement and support medical education techniques [18], particularly where the diagnostic reasoning aspect is concerned [15,19-22]. Notably, the combined use of NLP and ITS technologies in the simulation of virtual patients might promote students' learning by making the student-software interaction more similar to a real-life scenario, while simultaneously giving the student appropriate feedback after every simulated medical activity.

The primary aim of this study was to develop a VPS that combines interactions in natural language and ITS components, in order to set up a tool that would enable students to improve their clinical diagnostic reasoning skills. A secondary aim was to preliminarily assess the short-term potential changes in medical knowledge of a group of undergraduate students after the use of the VPS.

This article is structured with the Methods section describing the architecture and main development features of the program and with the Results section describing both the program's flow of use and the preliminary findings of a test performed on a population of undergraduate medical students.

Methods

The program we developed is named Hepius, after the Greek god of medicine, and it is structured to perform as both a VPS and an ITS.

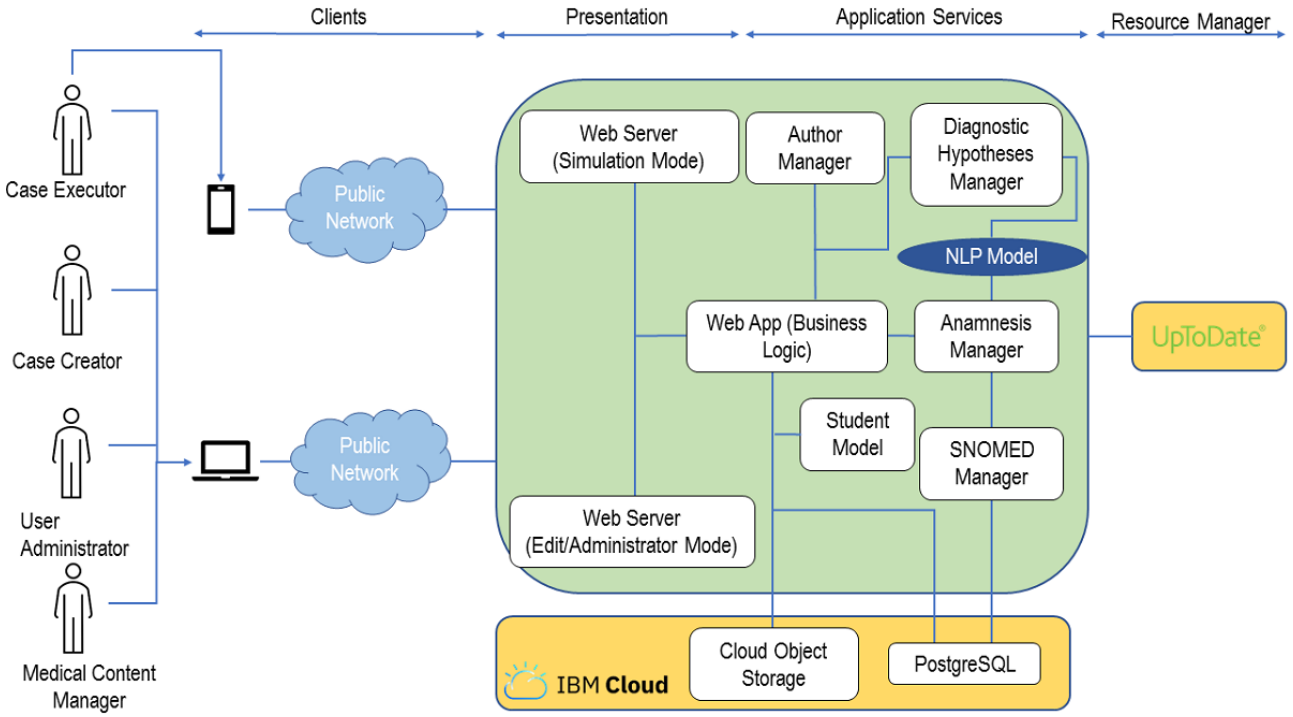
Program Architecture

The Hepius program architecture is outlined in Figure 1. Hepius has been designed and developed for four main categories of users: students, teachers, administrators, and medical content managers. The program is accessible through two main user interfaces: (1) a *mobile app*, developed using the Ionic Angular framework [23], that can be used to execute simulations and (2) a *web application*, developed using the PrimeFaces framework [24], that can be used to create and modify simulations or administer the system. Both user interface programs consume back-end services using representational state transfer application programming interfaces [25].

The Hepius back end has been developed according to the principles of microservices architecture [26] and it runs on the Cloud Foundry platform as a service (IBM Corp) [27]. The back-end components have been developed using three different programming languages: Java 8 (Oracle Corporation) as the main programming language, Python 3.7 (Python Software Foundation) for NLP services, and R 4.0 (The R Foundation) for the learner model.

The back end consumes an UpToDate service that is used to provide students with feedback. The Cloud Object Storage (IBM Corp) service is used as storage for multimedia files, whereas the PostgreSQL (Structured Query Language) (Compose) service is used as the main database. Both are provided in software-as-a-service mode by IBM Cloud.

Figure 1. Overview of the Hepius program architecture. NLP: natural language processing; SNOMED: Systematized Nomenclature of Medicine; SQL: Structured Query Language.



Natural Language Processing Algorithms

Interaction in natural language between the student and the program was developed for anamnesis, physical exams, medical test requests, and diagnostic hypothesis generation. Here we present, in detail, the diagnostic hypothesis generation and anamnesis modules.

Diagnostic Hypothesis Generation

When creating the simulation, the author decides which diagnostic hypotheses may be reasonable for the clinical case (ie, reference hypotheses). When the student formulates a diagnostic hypothesis in free text, Hepius assesses its correctness by calculating the Systematized Nomenclature of Medicine (SNOMED) graph path distance (ie, the minimum number of edges in any path connecting the two nodes) between the student's diagnostic hypothesis and all the reference hypotheses. If any of the reference hypotheses have zero distance from the student's hypothesis, then the student's hypothesis is marked as correct and is inserted into the differential diagnosis. Should the distance be greater than 5, the hypothesis is considered incorrect. Whenever the distance is between 1 and 4, the hypothesis is considered to be close to the correct one and the student is provided with feedback that points toward the closest reference hypothesis.

To find the best match between the input text string and the concepts in SNOMED ontology, we used Jaccard similarity [28] between token lists obtained from texts associated with concepts, including synonyms, after removal of stop words.

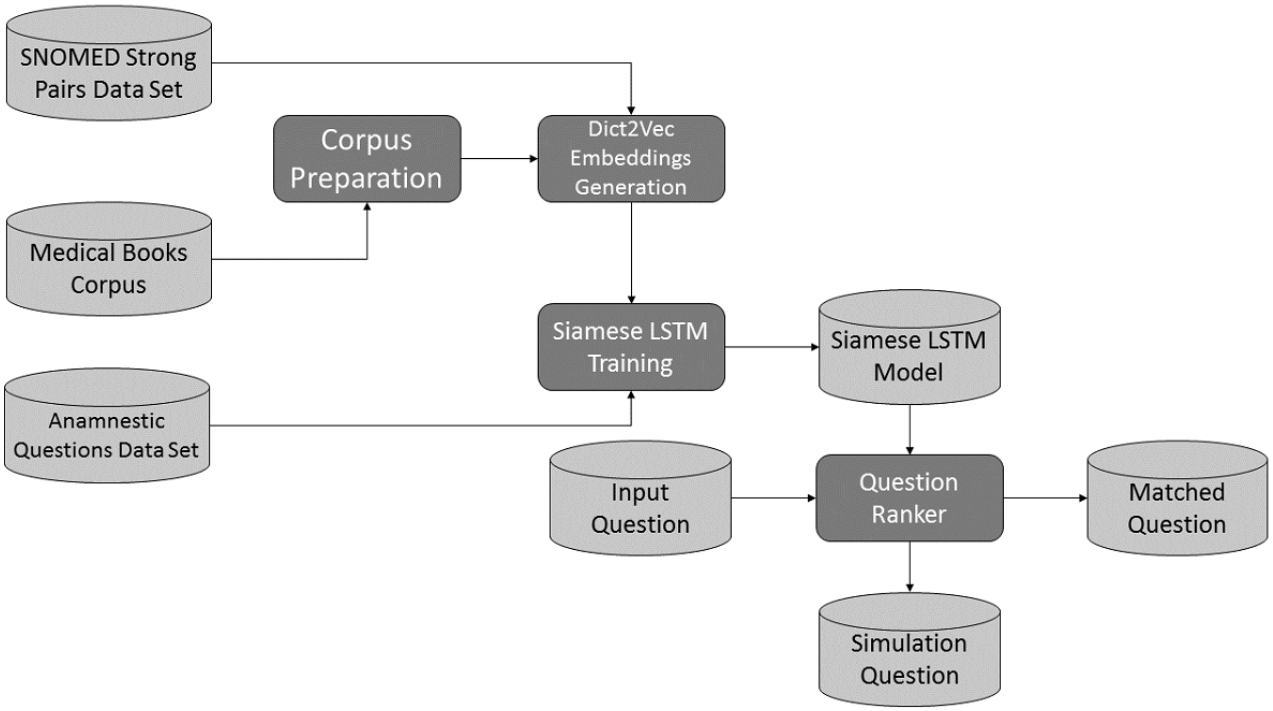
The entire diagnostic hypothesis module is implemented using only open-source code. The programming language is Python 3.7; the main libraries are Medical Terminologies for Python (PyMedTermino) [29], for interaction with the SNOMED CT (Clinical Terms) database, and Natural Language Toolkit (NLTK) 3.5 [30], for basic NLP operations (eg, tokenization).

Anamnesis

When the student formulates an anamnestic question, it is matched to the most semantically similar one present in the list of reference questions created by the teacher. The estimation of the semantic similarity of two sentences cannot simply be reduced to the semantic similarity of tokens inside the sentence (eg, using an ontology) because the meaning of a sentence depends on its extremely variable syntactic structure.

This *question matching* problem [31] has been addressed by developing an ad hoc pipeline of NLP algorithms (see Figure 2). The pipeline is based on a Siamese long short-term memory (SLSTM) network [32], trained on 7000 pairs of semantically equivalent and inequivalent anamnestic questions, that provides a probabilistic estimate of the semantic equivalence. This estimate is then used to rank all the reference anamnestic questions, thereby enabling the identification of the most similar one.

Figure 2. Pipeline of the history-taking natural language processing algorithms. Light grey cylinders identify data sources and dark grey boxes identify algorithms. LSTM: long short-term memory; SNOMED: Systematized Nomenclature of Medicine.



The SLSTM network requires a *word embedding* function [33] that converts words into tuples of real numbers (ie, vector representation) in such a way that semantically close words are transformed into vectors that are close according to a vector space metric [34]. Among the available unsupervised algorithms that learn word embedding, we decided to test Word2vec [35,36], Doc2vec [37], and fastText [38]. For all models, we generated our own embedding in an unsupervised way by means of the gensim library [39] using a corpus of medical textbooks and compared the overall pipeline performance with pretrained word embedding.

Using the medical textbooks corpus, the fastText word embedding that was generated proved to be superior in our setting compared to the other models, but it was still unable to correctly embed relevant pairs of medical synonyms. This problem has been addressed by the use of Dict2vec [40], introducing a form of weak supervision.

Long short-term memory (LSTM) networks [41] are neural networks that, like recurrent neural networks [41], can handle input sequences of arbitrary length by reusing at each computation step the same set of parameters, thereby reducing model complexity. LSTM networks are commonly used to tame the intrinsic instability of recurrent neural networks due to exploding and vanishing gradients [41]. Unlike more recent models, such as the Transformer [42], they are not designed for parallel computation being based on sequential inputs. In our context, we have two different inputs (ie, questions) that need to be compared; as a consequence, we need two LSTM networks that elaborate the inputs in parallel. For this purpose, we used SLSTM networks, whose key characteristic is that the two LSTM networks have exactly the same weights. The outputs of the networks are then compared using Manhattan distance [32].

The *question ranker* uses the trained SLSTM network model to compare the student input question with all the reference questions present in the simulation and ranks them according to the model output probability. A fixed probability threshold is used to decide whether the program should return a single question, multiple questions, or no questions. Returning multiple reference questions is undesirable because the program would be helping students in identifying reference questions that the student has not yet conceived, in contradiction with the didactical objective of having the student figure out the correct questions. On the other hand, returning matched questions only when the probability is very high could frustrate the students who would not receive correct semantic matches due to the fact that the algorithm has assigned low scores to these matches. The didactical decision we took was to fix a threshold and return all questions whose probability exceeds that threshold up to a maximum of three questions.

The anamnestic questions module is entirely written using open-source libraries to foster reproducibility. The programming language used to develop the module is Python 3.7. To generate the word embeddings, we used Dict2vec, for the reasons previously explained, by using the C code made available by the Dict2vect creators [43]. SLSTM networks were implemented using TensorFlow [44] and Keras [45]. The rationale underpinning the use of the SLSTM network is provided above; in addition, see Mueller and Thyagarajan [32] and Chen et al [46] for further details. An example of implementation strategy was found in Park [47]. The scikit-learn library [48] was used for basic data manipulations (eg, stratified train-test split). For basic NLP tasks (eg, tokenization and stemming), we used NLTK [29].

To test the above algorithms, we have developed six test sets, built out of six different simulations, with a total number of 547

questions, and measured the overall question matching accuracy. We obtained an accuracy greater than 70% for rank 1 matches and greater than 80% for rank 3 matches, as summarized in Table S1 of Multimedia Appendix 1.

Intelligent Tutoring System Development

ITSs are based on the concepts of an *inner loop* (ie, step-by-step feedback and hints during the execution of the learning unit) and an *outer loop* (ie, indications of what is the optimal next learning step) [49]. Out of the five key models of an ITS, in Hepius we implemented the following three: (1) the *domain model*, a decomposition of the knowledge corpus into concepts to be taught; (2) the *assessment model*, the definition of tests aimed at assessing the level of the student's understanding; and (3) the *learner model*, a mathematical model to predict learners' results when compared with assessments.

In Hepius, the *domain model* knowledge units are the diagnostic hypotheses (ie, diseases) and the diagnostic factors (ie, signs, symptoms, physical findings, and medical tests).

The Hepius *assessment model* works by comparing every student's action with the reference list containing all the possible correct actions written by the creator of the clinical case.

The Hepius *learner model* is a Bayesian Knowledge Tracing algorithm [50,51] that takes as an input the student performance in the execution of the binary analysis, for any diagnostic hypothesis, across multiple simulations. Bayesian Knowledge Tracing is based on a hidden Markov model (HMM) that provides an estimate of the probability that a student has a skill—in our context, the clinical understanding of a disease or diagnostic hypothesis—given his or her learning history—in our context, the results obtained during the analysis of the disease in previous simulations. To implement the algorithm, we used R packages HMM [52] and seqHMM [53].

Short-term Learning Test Protocol

A total of 15 medical students attending their fifth year at the Humanitas University Medical School in Italy participated in the test. Students were already acquainted with Hepius, as they had received specific introductory lectures and used them to perform simulated clinical cases in the preceding weeks.

The 2-hour-long test was conducted in the Humanitas University computer room, where students used individual desktop computers. On the day of the test, all students began by taking a uniform presimulation written test, made up of 22 multiple-choice questions (see Multimedia Appendix 2), to assess their baseline knowledge on chest pain and shortness of breath. The test topics had been previously covered during the semester. Each question was worth 1 point. Among the 22 questions, there were 11 *core* questions, presented in random order, which had been specifically designed to evaluate the knowledge that could be acquired directly by performing the simulation with Hepius. Thereafter, the students had 60 minutes to perform the simulation using the program. Notably, the chief complaints presented in the simulated clinical case were chest pain and dyspnea, with pulmonary embolism (PE) being the correct final diagnosis. Postsimulation, the students retook a multiple-choice question test, identical to the presimulation test, which was used to measure the changes in the number of right answers. Results were used as a proxy for the students' short-term knowledge acquisition. During the entire test period, students were not permitted to talk amongst themselves, consult written material, or use cell phones or similar devices. As shown in Multimedia Appendix 2, examples of core questions are questions 3 and 4. Given that the Hepius clinical case dealt with PE, question 3 was asking about the most common physical sign associated with PE (ie, tachycardia), whereas question 4 addressed the diagnostic relevance of low D-dimer plasma levels in excluding PE diagnosis, being that such a blood test was characterized by high negative predictive values. Both are crucial aspects of PE diagnosis and were addressed during the Hepius clinical case by expecting the student to look for these diagnostic factors when performing physical examinations and requesting medical tests, and to identify the correct relationship between these and the PE diagnostic hypothesis during the binary analysis. The remaining noncore questions dealt with issues presented and discussed during the semester's classes, as it was for PE, but not explicitly dealt with in the simulated clinical case. The aim of the noncore questions was to assess students' overall knowledge about the topics learned during half of the academic year; the aim was also to discriminate whether possible variations between pre- and postsimulation test scores were only related to knowledge that could be acquired through the simulated clinical case or, on the contrary, whether they were the result of a more generalized effect (eg, repeated-testing effect) [54,55].

Data are expressed as mean (SD). The Student *t* test for paired observations was used to evaluate, in each individual, the changes in the achieved scores before and after the simulation. Differences were considered significant at values of $P<.05$. Prism, version 8 (GraphPad Software), was used for statistical analyses.

Results

Overview

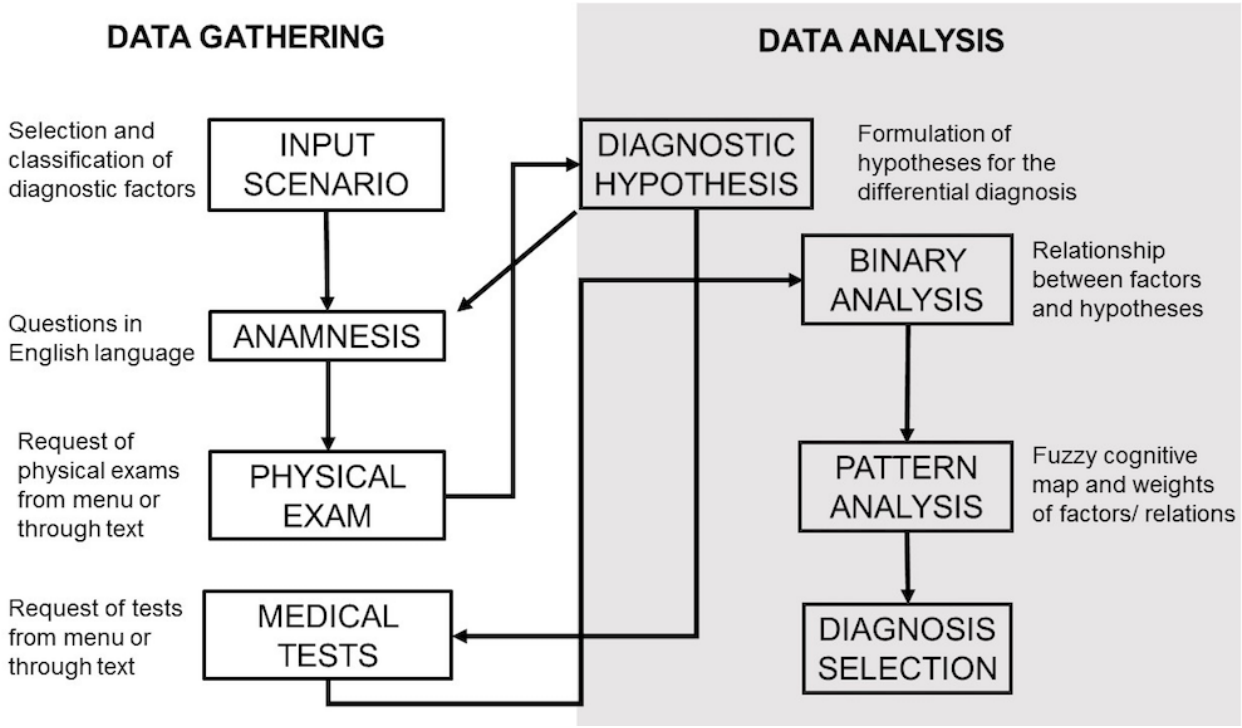
Hepius permits the creation of simulated clinical cases by human tutors and their execution by students. The creator of a simulated clinical case (ie, the tutor in charge) is responsible for creating a reference list containing all the clinically relevant information in the form of diagnostic factors (eg, body temperature = 39 °C), reasonable diagnostic hypotheses (eg, pneumonia and PE), the conceptual relationship between diagnostic factors and diagnostic hypotheses, and the correct final diagnosis. Further details on the creation of a simulation are provided in Multimedia Appendix 3.

Simulation of Clinical Cases With Interaction in Natural Language

The simulation of a clinical case with Hepius requires students to perform multiple actions that can be classified as either data gathering activities or data analysis activities (see Figure 3). *Data gathering* activities consist of obtaining diagnostic factors from the virtual patient through (1) examination of the patient's health records (ie, the input scenario), (2) anamnesis, (3) a physical exam, and (4) medical test requests. *Data analysis*

activities include (1) generating diagnostic hypotheses, (2) establishing causal links between diagnostic factors and diagnostic hypotheses (ie, binary analysis), and (3) estimating the magnitude of these links (ie, pattern analysis). Importantly, Hepius lets the student freely move back and forth within all sections of the simulation, allowing for clinical case reassessment.

Figure 3. Hepius' flow of use. The flowchart summarizes Hepius' structure and the diagnostic pathway that the student must follow to achieve the final diagnosis. Data gathering deals with the collection of anamnestic, physical, and instrumental data suitable for formulating likely diagnostic hypotheses. Data analysis refers to the differential diagnosis process. During data analysis, the student is asked to generate a diagnostic hypothesis by reasoning on the relationship between the gathered information and the single hypothesized diagnosis. This process is obtained by the binary analysis and the pattern analysis. This should train the learner to avoid ordering unnecessary tests. Selection of the final diagnosis ends the simulation.



In *data gathering* activities, the student has to collect all diagnostic factors that are potentially relevant for the final diagnosis. This is obtained by student-software interaction in natural language rather than by selecting a question or action from a predetermined list. The NLP algorithm then matches the student's anamnestic question with the most semantically similar reference question and provides its related answer. Natural language interaction is also available when a student performs the physical exam and asks for medical tests.

In the *data analysis* phase, the student works with the collected diagnostic factors to reach a final diagnosis. First, the student creates a differential diagnosis by writing her or his diagnostic hypotheses in natural language. Then, the NLP algorithm matches the student's diagnostic hypothesis to the semantically closest disease present in the SNOMED ontology. If the matched disease is present in the reference list, then the diagnostic hypothesis is considered correct and is included as part of the student's differential diagnosis. Once the student deems the differential diagnosis to be complete, the *binary analysis* can be performed (see Table 1). A table is automatically generated, listing the diagnostic factors (rows) and the diagnostic hypotheses (columns) identified thus far, in which the student is expected to outline whether each diagnostic factor increases, decreases, or does not affect the probability that the considered diagnostic hypothesis is the correct one.

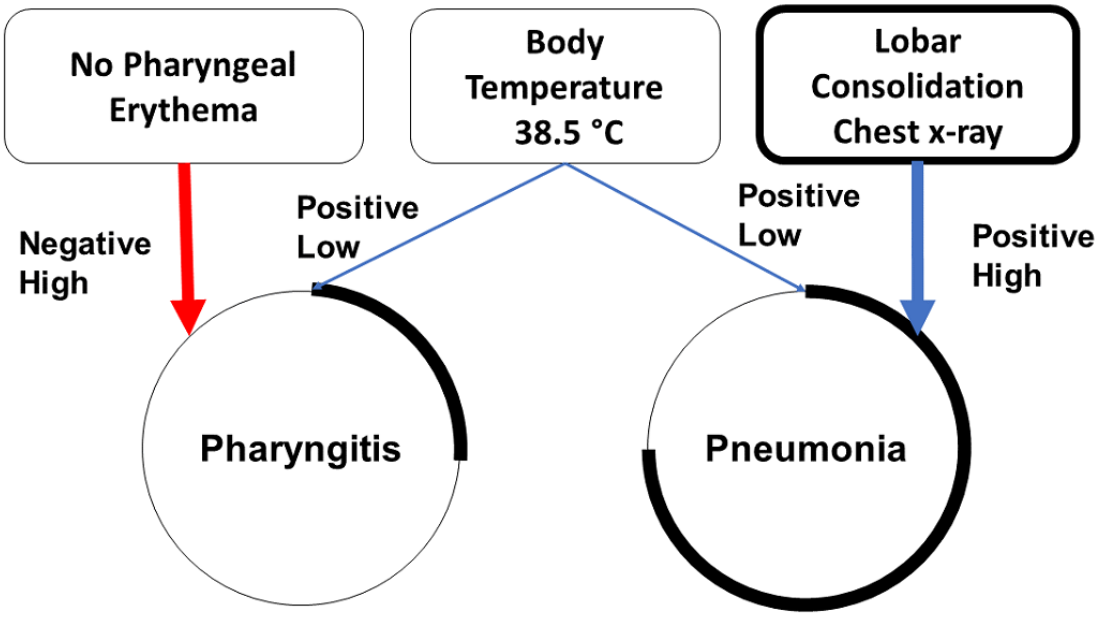
Table 1. Example of the binary analysis process.

Diagnostic factor[a] name	Diagnostic factor value	Diagnostic hypothesis[a]	
		Pharyngitis	Pneumonia
Body temperature	38.5 °C	I	I
Pharynx inspection	No pharyngeal erythema	D	N
Chest x-ray	Lobar consolidation	N	I

[a]The diagnostic factors and the diagnostic hypotheses are automatically added to rows and columns, respectively, for *binary analysis*. By selecting the boxes, the student actively chooses whether each diagnostic factor increases (I), decreases (D), or does not affect (N) the probability that the considered hypothesis will be the final diagnosis.

In the *pattern analysis*, the student can visualize and weigh the relationships among diagnostic factors and diagnostic hypotheses previously established during the *binary analysis*; see Figure 4 for further details. Once the student is satisfied with the analysis of the information previously gathered, the simulation can be ended by selecting the diagnostic hypothesis that is deemed to be correct.

Figure 4. Schematic overview of the pattern analysis process. Should the diagnostic factor increase the probability of the chosen diagnostic hypothesis, then the positive likelihood of such a relationship is represented by a connecting blue line. If a diagnostic factor is thought to decrease the likelihood of the diagnostic hypothesis, then the connecting line is depicted in red. When the diagnostic factor does not affect the diagnostic hypothesis, no connecting line is drawn. In addition, the student is asked to weigh the relevance of the diagnostic factors in relation to the hypothesized diagnoses. This is automatically translated into a graphic representation with an increase (positive) or decrease (negative) of the thickness of the connecting lines. In the example in the image, the presence of lobar consolidations on the chest x-ray was highly suggestive of pneumonia (positive high). Therefore, the thickness of the connecting line becomes wider. The circumference of the diagnostic hypothesis node was related to the probability that the chosen diagnosis was correct. As the probability of diagnosis increased, the portion of the highlighted circumference increased as well.



Intelligent Tutoring System

The ITS tracks all the student actions and provides real-time step-by-step feedback over the simulation's entire execution. For instance, if the student asks for a medical test that is absent in that clinical case reference list, he or she receives feedback stating that an inappropriate exam was asked for. As another example, should the diagnostic hypothesis made by the student (eg, pneumonia) be too general compared to the one in the reference list (eg, interstitial pneumonia), then feedback is given stating that the student should be more specific in generating the hypothesis. An exhaustive list of possible feedback is provided in Multimedia Appendix 4.

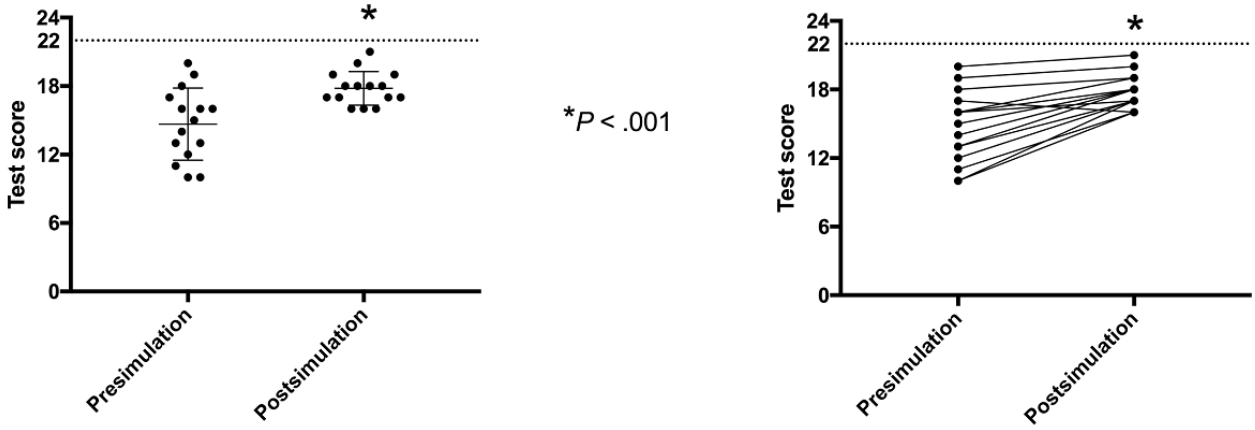
Furthermore, at the end of the simulation, the ITS provides feedback summarizing the diagnostic hypotheses in which the student has made more mistakes when addressing the binary analysis. In addition, links to the UpToDate topics related to these diagnostic hypotheses are given [56].

Moreover, the ITS logs all student actions, enabling post hoc learner analytics. In a related article currently under peer review [57], the possible applications of learner analytics are described in detail.

Short-term Learning Test Results

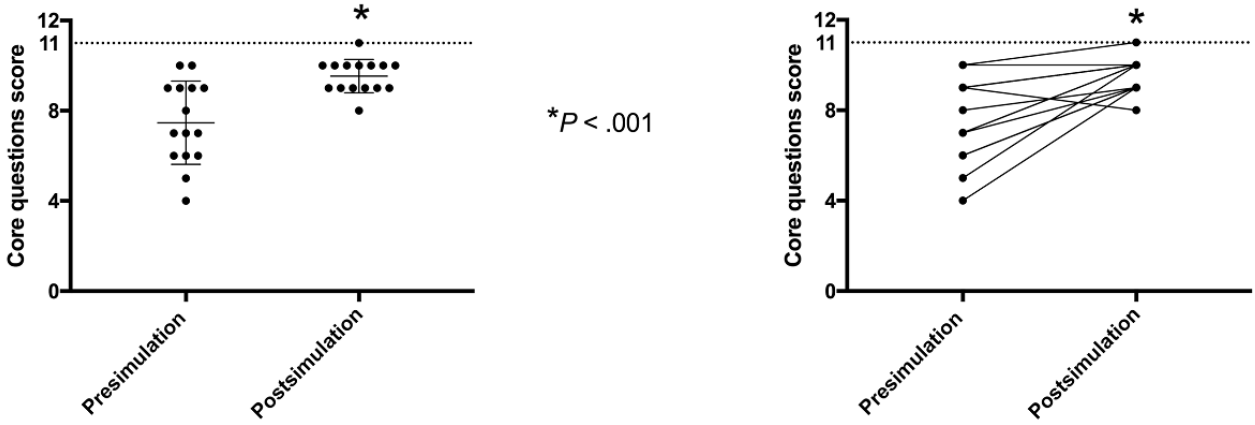
A significant improvement was found in the mean postsimulation overall test score compared to the presimulation overall test score (mean 17.8, SD 1.48, vs mean 14.6, SD 3.15, respectively; $P<.001$) (see Figure 5). Students' individual performances are shown in the right-hand graph of Figure 5. Only one subject's performance worsened after the simulation.

Figure 5. Overall pre- and postsimulation performance. Notice the significant improvement of the overall test score average after the use of Hepius (left-hand graph). Students' individual performances are shown in the right-hand graph.



There was a significant improvement in mean score for *core* questions from pre- to postsimulation (mean 7.46, SD 1.84, vs mean 9.53, SD 0.74, respectively; $P<.001$) (see Figure 6). Notably, out of the 15 students, 13 (87%) improved their *core* question scores from pre- to postsimulation. One student had no change and one obtained a lower score (see Figure 6, right-hand graph).

Figure 6. Pre- and postsimulation performance of core questions. The dashed horizontal line indicates the maximal reachable score. Scores are based on 15 students. A significant improvement in the mean score of core questions was observed from pre- to postsimulation tests (left-hand graph). Individual performances are displayed in the right-hand graph.



Discussion

In this paper, Hepius' most important features and the preliminary results obtained by its use in a medical undergraduate class are presented. Interaction in natural language and intelligent tutoring are the most important features of the program and are hereafter discussed.

Virtual Patient Simulators and Natural Language Processing

VPS may play an important role in medical education, particularly in training users in clinical diagnostic reasoning [58]. In the vast majority of VPSs, the interaction between the user and the simulated patient occurs by means of menus and the selection of predefined items [19,59]. The simulator recently developed by the New England Journal of Medicine Group [60] is such an example. It is aimed at training experienced doctors in facing COVID-19 cases that evolve over time according to the user's diagnostic and therapeutic interventions, which are selected from a predefined list of possibilities. Conversely, Hepius, which is specifically designed for undergraduate medical students, allows interaction through free text in the English language. We assumed that this type of automated interaction might better mirror real-life doctor-patient communication, thus increasing clinical simulation accuracy as previously suggested [22]. Furthermore, the absence of drop-down menus to select the most appropriate action highlights an important educational issue: students have to actively think about questions without getting hints by choosing prepackaged options. The same reasoning could be applied to diagnostic hypothesis generation.

Notably, a potential limitation of NLP techniques may be related to the low accuracy in interpreting questions. This can distract students from the focus of the task, as suggested in 2009 by Cook et al [10]. Nowadays, performance of the newest NLP algorithms has reached an accuracy as high as 95%, thus limiting

the risk of users' frustration for not having their questions understood by the simulator [22].

Intelligent Tutoring System

ITSs are programs aimed at providing immediate and customized instruction or feedback to learners, without interference from a human teacher [61]. These programs have been proven to be effective as teaching tools within different educational fields [12,13]. However, there are few studies about their use in the medical context. One of these is ReportTutor [62], which is an ITS aimed at helping pathology trainees to write correct biopsy reports in English natural language. Its tutoring activity stems from its capability to identify inaccuracies or missing features within the report and to give appropriate feedback to the trainees. Interestingly, ReportTutor shares NLP techniques with Hepius; however, those of ReportTutor are not devoted to mimicking the doctor-patient interaction.

Hepius integrates the key ITS concepts of *inner loop* (ie, step-by-step feedback and hints during the execution of the learning unit) and *outer loop* (ie, indications of what is the optimal next learning step) [49]. Inner loop feedback is given whenever a student performs an action. For example, if during the binary analysis the student wrongly states that the diagnostic factor *fever* decreases the likelihood of the patient having the diagnostic hypothesis *pneumonia*, then Hepius provides feedback indicating the correct relationship between these two factors. This type of feedback is important not only because it directly fosters learning but also because it allows students to complete their simulation, guiding them throughout the case. Outer loop feedback is instead given at the end of a simulation, according to the overall performance of the student. For example, if a user consistently makes mistakes in matching diagnostic factors to the diagnostic hypothesis *pneumonia*, the ITS recommends that the student review that specific topic by providing her or him with a link to the related UpToDate section. This type of automated feedback directly addresses weaknesses in the student's knowledge and provides him or her with suggestions on how to correct their mistakes.

Hepius as a Possible Didactical Tool for Clinical Diagnostic Reasoning

Hepius has been developed as a VPS with the aim of providing an automated training tool for clinical diagnostic reasoning. Clinical reasoning combines *intuitive thinking* (ie, heuristic thinking) and *analytical thinking*. Experienced doctors tend to apply heuristic thinking to an ordinary clinical case and revert to analytical thinking when the case is rare or complex. On the other hand, less experienced physicians mainly rely on analytical thinking [63].

Hepius has been developed to target undergraduate medical students in order to train them in analytical thinking. This mental process is applied, for instance, during the binary analysis, where the student is asked to disclose the causal relationship between each single diagnostic factor and diagnostic hypothesis. In addition, through the pattern analysis, Hepius provides the student with the possibility of visually addressing the relationships between diseases and clinical findings, in a process similar to conceptual maps [64]. Overall, these analytical exercises are expected to help students enhance their diagnostic skills and medical knowledge, although no robust evidence is presently available, except for our preliminary findings. These shall be briefly discussed below.

The capability of Hepius to enhance medical knowledge in the short term was preliminarily evaluated among 15 students attending their fifth year at the Humanitas University Medical School. They completed an identical test, composed of multiple-choice questions, before and after the clinical case simulation by Hepius. We hypothesized that, in such a way, the test would provide proper insight into the potential changes in students' knowledge on the specific issue dealt with during the simulation (ie, PE). In keeping with previous reports highlighting the educational capabilities of VPSs [10,14], in this study, Hepius use resulted in an increase in the performance scores of almost all the students. This was the case for the students who had good baseline performance as well as for those whose initial performance was poor. Taken together, these findings suggest that, in the short term, Hepius might act as a didactical tool.

However, in spite of its promising features, it is important to stress that Hepius cannot fully replace a skilled human tutor working one on one with a learner [65]. Instead, in keeping with a *blended* approach, it is intended to be used as a classroom assistant as well as a tool for distance learning. Indeed, as with any VPS, Hepius allows for proper social distancing; therefore, it is potentially useful in overcoming the didactical problem regarding the temporary inability to attend clinical facilities in the setting of the COVID-19 outbreak.

Limitations

As with any automated didactical tool, students' performance using Hepius is characterized by a learning curve, and its optimized use requires initial tutoring. This is presently provided via a video tutorial and should be refined by teachers through ad hoc online lectures, in accordance with the concept of orchestration of intelligent learning environments [15,66].

Accuracy of the diagnostic hypothesis generation module has not been estimated due to the lack of a comprehensive test set. Also, we have not attempted to use language modeling or semantic similarity algorithms based on a deep learning algorithm approach. Both activities are objectives for future work. Finally, the short-term learning test has been carried out among a small number of students and using a limited pool of questions. Thus, our findings should be regarded as preliminary results that must be confirmed in future studies and further validated on larger cohorts.

Conclusions

Shortage of human resources, increasing educational costs, and the need to keep social distances in response to the COVID-19 worldwide outbreak have prompted the necessity of automated clinical training methods designed for distance learning. We have developed a VPS named Hepius that, by natural language interaction and an ITS component, might help students to improve their clinical diagnostic reasoning skills without necessarily requiring the presence of human tutors or the need for the student to be at the bedside of a real patient.

Implementation of additional features, such as therapy and patient management modules, can be pursued to make Hepius suitable for application in postgraduate residency programs and continuing medical education.

As a preliminary assessment of its educational impact, we found that the use of Hepius may enhance students' short-term knowledge. Ad hoc studies using larger populations are needed to confirm this result and to investigate Hepius' actual long-term didactical capability.

Acknowledgments

We are thankful to Giorgio Ferrari, Chief Executive Officer (CEO) of Humanitas University; Luciano Ravera, CEO of Humanitas Research Hospital; and Fabrizio Renzi, IBM Italy Director of Technology and Innovation, for their initial and continuous support. Key stakeholders include Alessandra Orlandi, Chief Innovation Officer of Humanitas; Elena Sini, Chief Information Officer of Humanitas Research Hospital; Victor Saveski, Chief Web and Social of Humanitas; Valeria Ingrosso, Humanitas Special Programs; and Giovanna Camorali, Business Development Executive at IBM. The functional requirements team includes Anna Giulia Bottaccioli, Vita-Salute San Raffaele University, Milan, Italy. The development team includes Luca Vinciotti, Database Architect at IBM Italy; Michele Savoldelli, Back-End Architect at IBM Italy; Jacopo Balocco, Learner Model Developer at IBM Italy; and Valerio Chieppa, Pattern Analysis Developer at IBM Italy. We acknowledge the contribution of Marco Asti from UpToDate, as the evidence-based knowledge reference. This study was cofunded by Humanitas Clinical and Research Center, Humanitas University, and IBM Italy.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Natural language matchers.
[DOC File , 35 KB - medinform_v9i4e24073_app1.doc]

Multimedia Appendix 2
Multiple-choice question test.
[DOCX File , 18 KB - medinform_v9i4e24073_app2.docx]

Multimedia Appendix 3
Creation of a simulation.
[DOC File , 26 KB - medinform_v9i4e24073_app3.doc]

Multimedia Appendix 4
Inner-loop feedback.
[DOCX File , 17 KB - medinform_v9i4e24073_app4.docx]

References

1. Andel C, Davidow SL, Hollander M, Moreno DA. The economics of health care quality and medical errors. J Health Care Finance 2012;39(1):39-50. [Medline: 23155743]
2. Schmidt HG, Mamede S. How to improve the teaching of clinical reasoning: A narrative review and a proposal. Med Educ 2015 Oct;49(10):961-973. [doi: 10.1111/medu.12775] [Medline: 26383068]
3. Ramani S, Leinster S. AMEE Guide no. 34: Teaching in the clinical environment. Med Teach 2008;30(4):347-364. [doi: 10.1080/01421590802061613] [Medline: 18569655]
4. Lucey CR, Johnston SC. The transformational effects of COVID-19 on medical education. JAMA 2020 Sep 15;324(11):1033-1034. [doi: 10.1001/jama.2020.14136] [Medline: 32857137]
5. Rose S. Medical student education in the time of COVID-19. JAMA 2020 Jun 02;323(21):2131-2132. [doi: 10.1001/jama.2020.5227] [Medline: 32232420]
6. Al-Balas M, Al-Balas HI, Jaber HM, Obeidat K, Al-Balas H, Aborajooh EA, et al. Correction to: Distance learning in clinical medical education amid COVID-19 pandemic in Jordan: Current situation, challenges, and perspectives. BMC Med Educ 2020 Dec 16;20(1):513 [FREE Full text] [doi: 10.1186/s12909-020-02428-3] [Medline: 33327927]
7. Wayne DB, Green M, Neilson EG. Medical education in the time of COVID-19. Sci Adv 2020 Jul;6(31):eabc7110 [FREE Full text] [doi: 10.1126/sciadv.abc7110] [Medline: 32789183]
8. Julian D, Smith R. Developing an intelligent tutoring system for robotic-assisted surgery instruction. Int J Med Robot 2019 Dec;15(6):e2037. [doi: 10.1002/rcs.2037] [Medline: 31509636]

9. Posel N, Mcgee JB, Fleiszer DM. Twelve tips to support the development of clinical reasoning skills using virtual patient cases. Med Teach 2014 Dec 19;37(9):813-818. [doi: 10.3109/0142159x.2014.993951]
10. Cook D, Triola MM. Virtual patients: A critical literature review and proposed next steps. Med Educ 2009 Apr;43(4):303-311. [doi: 10.1111/j.1365-2923.2008.03286.x] [Medline: 19335571]
11. Berman NB, Durning SJ, Fischer MR, Huwendiek S, Triola MM. The role for virtual patients in the future of medical education. Acad Med 2016;91(9):1217-1222. [doi: 10.1097/acm.0000000000001146]
12. Mousavinasab E, Zarifsanaiey N, Niakan Kalhori SR, Rakhshan M, Keikha L, Ghazi Saeedi M. Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. Interac Learn Environ 2018 Dec 18;29(1):142-163. [doi: 10.1080/10494820.2018.1558257]
13. Kulik JA, Fletcher JD. Effectiveness of intelligent tutoring systems. Rev Educ Res 2016 Mar;86(1):42-78. [doi: 10.3102/0034654315581420]
14. Vanlehn K, Lynch C, Schulze K, Shapiro JA, Shelby R, Taylor L, et al. The Andes physics tutoring system: Five years of evaluations. In: Proceedings of the 12th International Conference on Artificial Intelligence in Education. 2005 Presented at: 12th International Conference on Artificial Intelligence in Education; July 18-22, 2005; Amsterdam, the Netherlands p. 678-685 URL: https://people.engr.ncsu.edu/cflynch/Papers/VanLehn_andes_aied2005_final.pdf
15. du Boulay B. Escape from the Skinner Box: The case for contemporary intelligent learning environments. Br J Educ Technol 2019 Jul 15;50(6):2902-2919. [doi: 10.1111/bjet.12860]
16. Sundararajan SC, Nitta SV. Designing engaging intelligent tutoring systems in an age of cognitive computing. IBM J Res Dev 2015 Nov;59(6):10:1-10:9. [doi: 10.1147/jrd.2015.2464085]
17. Daniel M, Rencic J, Durning SJ, Holmboe E, Santen SA, Lang V, et al. Clinical reasoning assessment methods. Acad Med 2019;94(6):902-912. [doi: 10.1097/acm.0000000000002618]
18. Afzal S, Dhamecha TI, Gagnon P, Nayak A, Shah A, Carlstedt-Duke J, et al. AI medical school tutor: Modelling and implementation. In: Proceedings of the International Conference on Artificial Intelligence in Medicine. Cham, Switzerland: Springer International Publishing; 2020 Presented at: International Conference on Artificial Intelligence in Medicine; August 25-28, 2020; Minneapolis, MN p. 133-145. [doi: 10.1007/978-3-030-59137-3_13]
19. Sapci AH, Sapci HA. Artificial intelligence education and tools for medical and health informatics students: Systematic review. JMIR Med Educ 2020 Jun 30;6(1):e19285 [FREE Full text] [doi: 10.2196/19285] [Medline: 32602844]
20. Chary M, Parikh S, Manini A, Boyer E, Radeos M. A review of natural language processing in medical education. West J Emerg Med 2019 Jan;20(1):78-86 [FREE Full text] [doi: 10.5811/westjem.2018.11.39725] [Medline: 30643605]
21. Khaled D. Natural language processing and its use in education. Int J Adv Comput Sci Appl 2014;5(12):72-76 [FREE Full text] [doi: 10.14569/ijacsa.2014.051210]
22. Persad A, Stroulia E, Forgie S. A novel approach to virtual patient simulation using natural language processing. Med Educ 2016 Nov;50(11):1162-1163. [doi: 10.1111/medu.13197] [Medline: 27762013]
23. Ionic Angular overview. Ionic. 2019. URL: https://ionicframework.com/docs/angular/overview [accessed 2021-03-19]
24. PrimeFaces. URL: https://www.primefaces.org [accessed 2021-03-19]
25. What is a REST API? Red Hat. URL: https://www.redhat.com/en/topics/api/what-is-a-rest-api [accessed 2021-03-19]
26. What are microservices? Red Hat. URL: https://www.redhat.com/en/topics/microservices/what-are-microservices [accessed 2021-03-19]
27. IBM Cloud Foundry. IBM. URL: https://www.ibm.com/cloud/cloud-foundry [accessed 2021-03-19]
28. Jaccard index. Wikipedia. URL: https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=977019056 [accessed 2021-03-19]
29. PyMedTermino. Python Package Index. URL: https://pypi.org/project/PyMedTermino/ [accessed 2021-03-19]
30. Natural Language Toolkit. URL: https://www.nltk.org/ [accessed 2021-03-19]
31. Wu G, Lan M. Exploring traditional method and deep learning method for question retrieval and answer ranking in community question answering. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 2016 Presented at: 10th International Workshop on Semantic Evaluation (SemEval-2016); June 16-17, 2016; San Diego, CA p. 872-878 URL: https://www.aclweb.org/anthology/S16-1135.pdf [doi: 10.18653/v1/s16-1135]
32. Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16). 2016 Presented at: Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16); February 12-17, 2016; Phoenix, AZ p. 2786-2792 URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/12195/12023
33. Goldberg Y, Levy O. word2vec explained: Deriving Mikolov et al's negative-sampling word-embedding method. ArXiv. Preprint posted online on February 15, 2014. [FREE Full text]
34. Eisenstein J. Introduction to Natural Language Processing. Cambridge, MA: The MIT Press; 2019.
35. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations (ICLR 2013). 2013 Presented at: 1st International Conference on Learning Representations (ICLR 2013); May 2-4, 2013; Scottsdale, AZ URL: https://arxiv.org/pdf/1301.3781
36. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'13).

2013 Presented at: 27th International Conference on Neural Information Processing Systems (NIPS'13); December 5-10, 2013; Lake Tahoe, NV p. 3111-3119 URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf
37. Le Q, Mikolov T. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. 2014 Presented at: 31st International Conference on Machine Learning; June 21-26, 2014; Beijing, China p. 1188-1196 URL: http://proceedings.mlr.press/v32/le14.pdf
38. Bojanowski P, Grave E, Joulin A, Mikolov Y. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017;5:135-146 [FREE Full text] [doi: 10.1162/tacl_a_00051]
39. gensim. GitHub. URL: https://github.com/RaRe-Technologies/gensim [accessed 2021-03-19]
40. Tissier J, Gravier C, Habrard A. Dict2vec : Learning word embeddings using lexical dictionaries. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; September 9-11, 2017; Copenhagen, Denmark p. 254-263 URL: https://www.aclweb.org/anthology/D17-1024.pdf [doi: 10.18653/v1/d17-1024]
41. Aggarwal CC. Neural Networks and Deep Learning. Cham, Switzerland: Springer International Publishing; 2018.
42. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017). 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA URL: https://arxiv.org/pdf/1706.03762.pdf
43. dict2vec. GitHub. URL: https://github.com/tca19/dict2vec [accessed 2021-03-19]
44. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. Google Research. 2015. URL: https://research.google/pubs/pub45166.pdf [accessed 2021-03-19]
45. Keras. URL: https://keras.io [accessed 2021-03-19]
46. Chen Z, Zhang H, Zhang X, Zhao L. Quora Question Pairs. static.hongbozhang.me. URL: http://static.hongbozhang.me/doc/Quora.pdf [accessed 2021-03-19]
47. Park SK. Siamese-LSTM. GitHub. URL: https://github.com/likejazz/Siamese-LSTM#readme [accessed 2021-03-19]
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
49. Essa A. A possible future for next generation adaptive learning systems. Smart Learn Environ 2016;3:1-24 [FREE Full text] [doi: 10.1186/s40561-016-0038-y]
50. Corbett AT, Anderson JR. Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model User-adapt Interact 1994;4:253-278. [doi: 10.1007/bf01099821]
51. Yudelson MV, Koedinger KR, Gordon GJ. Individualized Bayesian knowledge tracing models. In: Proceedings of the International Conference on Artificial Intelligence in Education. 2013 Presented at: International Conference on Artificial Intelligence in Education; July 9-13, 2013; Memphis, TN p. 171-180. [doi: 10.1007/978-3-642-39112-5_18]
52. Helske S, Helske J. Mixture hidden Markov models for sequence data: The seqHMM package in R. J Stat Softw 2019;88(3):1-32 [FREE Full text] [doi: 10.18637/jss.v088.i03]
53. Helske J, Helske S. Package ‘seqHMM’: Mixture hidden Markov models for social sequence data and other multivariate, multichannel categorical time series. The Comprehensive R Archive Network. 2019 Oct. URL: https://cran.r-project.org/web/packages/seqHMM/seqHMM.pdf [accessed 2021-03-19]
54. Schuelper N, Ludwig S, Anders S, Raupach T. The impact of medical students' individual teaching format choice on the learning outcome related to clinical reasoning. JMIR Med Educ 2019 Jul 22;5(2):e13386 [FREE Full text] [doi: 10.2196/13386] [Medline: 31333193]
55. Roediger HL, Karpicke JD. The power of testing memory: Basic research and implications for educational practice. Perspect Psychol Sci 2006 Sep;1(3):181-210. [doi: 10.1111/j.1745-6916.2006.00012.x] [Medline: 26151629]
56. UpToDate. URL: https://www.uptodate.com [accessed 2019-11-28]
57. Furlan R, Gatti M, Menè R, Shiffer D, Marchiori C, Levra AG, et al. Learning analytics applied to clinical diagnostic reasoning using an NLP-based virtual patient simulator: A case study. JMIR Preprints. Preprint posted online on September 16, 2020. [FREE Full text] [doi: 10.2196/preprints.24372]
58. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual patient simulations in health professions education: Systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res 2019 Jul 02;21(7):e14676 [FREE Full text] [doi: 10.2196/14676] [Medline: 31267981]
59. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: A systematic review and meta-analysis. Acad Med 2010;85(10):1589-1602. [doi: 10.1097/acm.0b013e3181edfe13]
60. Abdulnour RE, Lieber J, Faselis C, Sternschein R, Hayden RM, Massaro A, et al. Covid-19 Rx: Treatment simulations. NEJM Group. 2020. URL: https://covid19rx.nejm.org/landing/index.html [accessed 2021-03-19]
61. Almasri A, Ahmed A, Al-Masri N, Abu Sultan Y, Mahmoud AY, Zaqout I, et al. Intelligent tutoring systems survey for the period 2000-2018. Int J Acad Eng Res 2019 May;3(5):21-37 [FREE Full text]

62. El Saadawi GM, Tseytlin E, Legowski E, Jukic D, Castine M, Fine J, et al. A natural language intelligent tutoring system for training pathologists: Implementation and evaluation. Adv Health Sci Educ Theory Pract 2008 Dec;13(5):709-722 [FREE Full text] [doi: 10.1007/s10459-007-9081-3] [Medline: 17934789]
63. Croskerry P. Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. Adv Health Sci Educ Theory Pract 2009 Sep;14 Suppl 1:27-35. [doi: 10.1007/s10459-009-9182-2] [Medline: 19669918]
64. Daley B, Torre DM. Concept maps in medical education: An analytical literature review. Med Educ 2010 May;44(5):440-448. [doi: 10.1111/j.1365-2923.2010.03628.x] [Medline: 20374475]
65. du Boulay B. Artificial intelligence as an effective classroom assistant. IEEE Intell Syst 2016 Nov;31(6):76-81. [doi: 10.1109/mis.2016.93]
66. Dillenbourg P. Design for classroom orchestration. Comput Educ 2013 Nov;69:485-492. [doi: 10.1016/j.compedu.2013.04.013]

Abbreviations

CEO: Chief Executive Officer
HMM: hidden Markov model
ITS: intelligent tutoring system
LSTM: long short-term memory
NLP: natural language processing
NLTK: Natural Language Toolkit
PE: pulmonary embolism
PyMedTermino: Medical Terminologies for Python
SLSTM: Siamese long short-term memory
SNOMED: Systematized Nomenclature of Medicine
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms
SQL: Structured Query Language
VPS: virtual patient simulator

Edited by G Eysenbach; submitted 03.09.20; peer-reviewed by M Bruno, V Franzoni; comments to author 18.11.20; revised version received 22.12.20; accepted 25.02.21; published 09.04.21.

Please cite as:
Furlan R, Gatti M, Menè R, Shiffer D, Marchiori C, Giaj Levra A, Saturnino V, Brunetta E, Dipaola F
A Natural Language Processing–Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study
JMIR Med Inform 2021;9(4):e24073
URL: https://medinform.jmir.org/2021/4/e24073
doi:10.2196/24073
PMID:33720840

Original Paper

Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning

Haishuai Wang[1,2], PhD; Paul Avillach[1], MD, PhD

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, United States

[2]Department of Computer Science and Engineering, Fairfield University, Fairfield, CT, United States

Corresponding Author:
Paul Avillach, MD, PhD
Department of Biomedical Informatics
Harvard Medical School
10 Shattuck Street
Boston, MA, 02115
United States
Phone: 1 617 432 2144
Email: Paul_Avillach@hms.harvard.edu

Abstract

Background: In the United States, about 3 million people have autism spectrum disorder (ASD), and around 1 out of 59 children are diagnosed with ASD. People with ASD have characteristic social communication deficits and repetitive behaviors. The causes of this disorder remain unknown; however, in up to 25% of cases, a genetic cause can be identified. Detecting ASD as early as possible is desirable because early detection of ASD enables timely interventions in children with ASD. Identification of ASD based on objective pathogenic mutation screening is the major first step toward early intervention and effective treatment of affected children.

Objective: Recent investigation interrogated genomics data for detecting and treating autism disorders, in addition to the conventional clinical interview as a diagnostic test. Since deep neural networks perform better than shallow machine learning models on complex and high-dimensional data, in this study, we sought to apply deep learning to genetic data obtained across thousands of simplex families at risk for ASD to identify contributory mutations and to create an advanced diagnostic classifier for autism screening.

Methods: After preprocessing the genomics data from the Simons Simplex Collection, we extracted top ranking common variants that may be protective or pathogenic for autism based on a chi-square test. A convolutional neural network–based diagnostic classifier was then designed using the identified significant common variants to predict autism. The performance was then compared with shallow machine learning–based classifiers and randomly selected common variants.

Results: The selected contributory common variants were significantly enriched in chromosome X while chromosome Y was also discriminatory in determining the identification of autistic individuals from nonautistic individuals. The *ARSD*, *MAGEB16*, and *MXRA5* genes had the largest effect in the contributory variants. Thus, screening algorithms were adapted to include these common variants. The deep learning model yielded an area under the receiver operating characteristic curve of 0.955 and an accuracy of 88% for identifying autistic individuals from nonautistic individuals. Our classifier demonstrated a considerable improvement of ~13% in terms of classification accuracy compared to standard autism screening tools.

Conclusions: Common variants are informative for autism identification. Our findings also suggest that the deep learning process is a reliable method for distinguishing the diseased group from the control group based on the common variants of autism.

(JMIR Med Inform 2021;9(4):e24754) doi:10.2196/24754

KEYWORDS

deep learning; autism spectrum disorder; common genetic variants, diagnostic classification

Introduction

Autism spectrum disorder (ASD) is a common neurodevelopmental disorder that begins early in childhood and lasts throughout a person's life. In the United States, around 1 out of 59 children have been diagnosed with ASD. People with ASD have characteristic social communication deficits and repetitive behaviors. Early detection of ASD enables timely interventions for children with ASD. Such interventions could provide the best opportunity to improve outcomes as opposed to treatments started after diagnosis. The epigenetic landscape has revealed that ASD may result from a complex regulatory network, including epigenetic, genetic, and environmental factors [1]. Although the causes of ASD remain unknown, recent studies have found that ASDs are 80% reliant on the inherited genes [1-3]. Twin studies of ASD show heritability as a highly responsible factor causing the disorder [4,5]. Therefore, identifying genomic mutations for autism based upon genotype information for early diagnosis of autism is significantly important. The genetic landscape of ASD is heterogeneous and consists of various types of genetic abnormalities involving almost all genes (eg, *SHANK3, SHANK2, CHD8, SEMA5A, DOCK4*) with different levels of penetrance [6-8]. Thus, autism studies have been conducted with different types of genetic variants [9-14], including *de novo* or inherited copy number variants, multiple hits, rare variants, common variants, and genetic pathways associated with ASD.

Rare variants, both inherited and *de novo*, are causal in 10%-30% of people with ASDs [15-17]. Although risk-associated genes of autism have been identified from rare variations, recent studies have shown that most genetic risks for ASD reside with common variations [18]. A Population-Based Autism Genetics and Environment Study on a Swedish epidemiological sample shows synthesis of results regarding the genetic architecture of ASD and concludes that inherited rare variations constitute a smaller fraction of the total heritability than common variations [18]. Several genome-wide association studies have also examined that 15%-40% of the genetic risk associated with ASD diagnosis is tagged by common variants [19-21]. Therefore, common variants may be informative with respect to the identification of ASD. Numerous studies have since used genetic information to predict the diagnosis of ASD. A single nucleotide polymorphism–based test has been demonstrated to allow for early identification of ASD [22]. In this study, they applied machine learning to identify single nucleotide polymorphisms to generate a predictive classifier for ASD diagnosis and have proved and concluded that the predictive classifier can be a tool to estimate the probability of at-risk status for ASD. To enable earlier and more accurate diagnoses of ASD, a statistical model has been developed for autism to analyze measurements of metabolite concentrations and it indicated that the metabolites under consideration are highly associated with an autism diagnosis [23]. A gene expression–based study has demonstrated that the accuracy of distinguishing ASD subgroups from nonautistic controls by using a support vector machine can be up to 94% [24]. Combining a brain-specific gene network with a complementary machine learning approach has also been conducted to present a genome-wide prediction of autism risk genes [25]. However, none of the existing works provide adequate accuracy or specificity that can be used for autism diagnosis with common variations. Recently, deep neural networks have achieved record-breaking performance in a variety of real-world applications [26-29]. In this study, we adapt deep learning to the task of predicting ASD and propose a deep learning–based framework, named DeepAutism, to predict autism disorder phenotypes by using common variants.

This study first identified significant common variants that may be protective or pathogenic for ASD as well as their additive contribution to ASD; therefore, deep learning models are applicable using common variants. Then, this study applied deep learning prediction algorithms to verify the identified common variants and generate a predictive classifier for ASD diagnosis. The results were tested on a hold-out test data set from the Simons Simplex Collection (SSC), and the proposed strategic approach achieved the best performance in distinguishing the diseased group from the control group based on selected significant common variants of ASD.

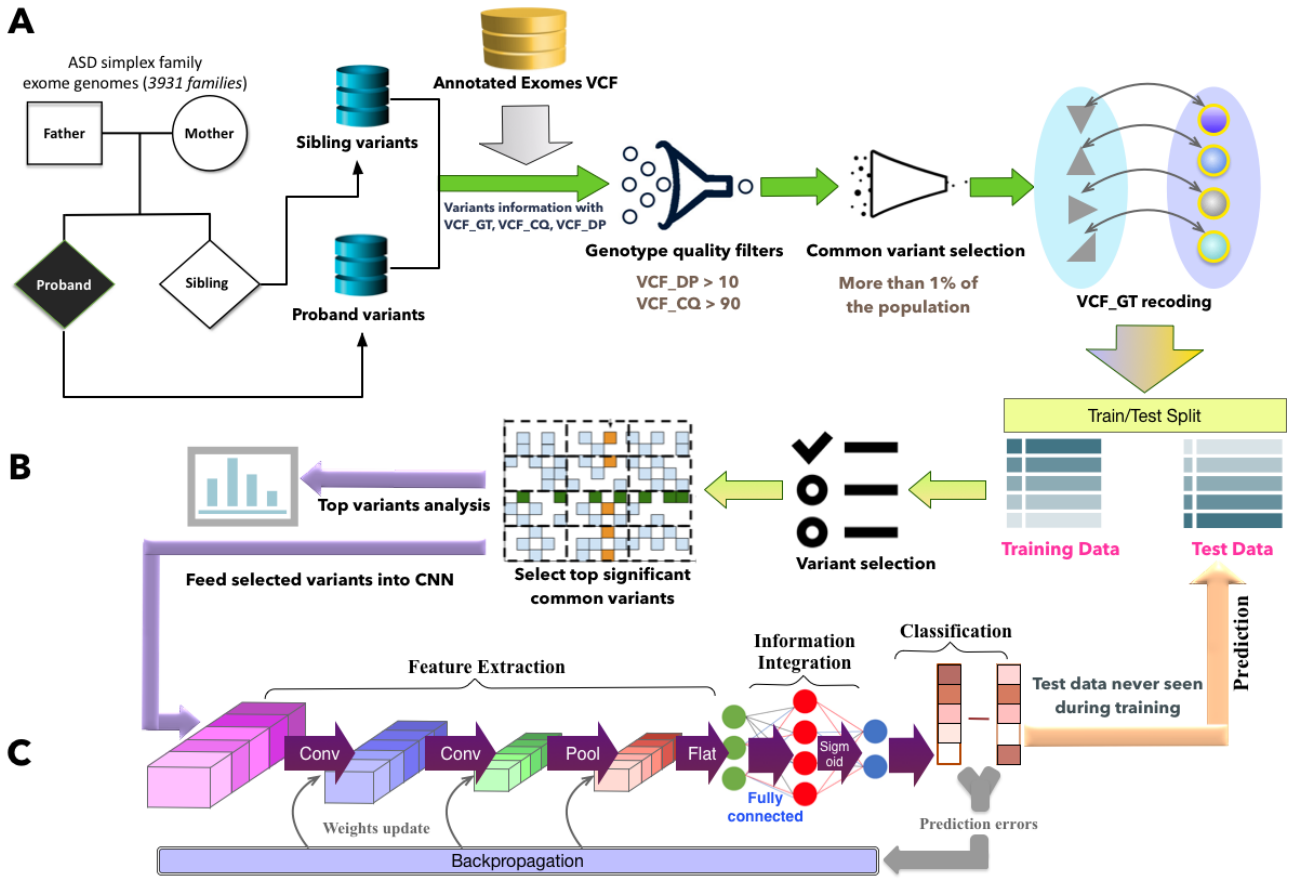
The objectives of this study were to (1) discover significant common variants that may be protective or pathogenic for ASD, (2) create an advanced diagnostic classifier for autism screening based on the identified common variants, and (3) verify the developed classifier and significant common variants across thousands of simplex families.

Methods

Data Set

We used an autism data set from the SSC [30]. The SSC data consist of 2600 simplex families, each of which has 1 child affected with ASD (a proband), unaffected parents, and at least one unaffected sibling. The data consist of 3931 individuals whose exome sequences are available (Figure 1A), and 2249 samples of these individuals are labeled as diseased group (ASD). From the SSC data set, we can query the specific variables for exome variants (Figure 1A), and the variants are in the variant call format (VCF). There are more than 1.5 million variants in the data set, which has the genotype information along with read depth, allele depth, and genotype quality. In the VCF data, VCF_GT represents the genotype quality, encoded as allele values separated by "/," such as "0/1" and "2/3", where 0 represents the reference allele, 1 for the first allele listed in the alternate allele, 2 for the second allele listed in the alternate allele, and so on. Thus, VCF_GT can be "0/0", "0/1", "2/0", "1/2", and so on. The read depth is denoted as VCF_DP, and the conditional genotype quality is denoted as VCF_CQ. We mainly used the information of VCF_GT, VCF_DP, and VCF_CQ in the SSC data for this study. The Harvard Medical School Research Ethics Committee approved this study.

Figure 1. Overall framework for deciphering contributory common variants and predicting autism spectrum disorder diagnosis. A. Data preprocessing. VCF_GT recoding is to encode VCF_GT values as dummy variables. If both alleles are reference alleles, it is encoded as 0; if both alleles are alternate alleles, it is encoded as 2; otherwise, it is 1. B. Data split and significant variant selection. The data set was split into training set and test set. Variants were ranked based on their chi-score and *P* value, and only top ranked (high chi-score value and low *P* value) variants were selected as contributory common variants for autism spectrum disorder. C. Convolutional neural network classifier. The selected significant common variants in the training data were fed into a convolutional neural network to train a classifier. Thereafter, the trained model was applied on the test data for autism spectrum disorder diagnosis prediction. ASD: autism spectrum disorder; CNN: convolutional neural network; SSC: Simons Simplex Collection; VCF: variant call format; VCF_CQ: variant call format-conditional genotype quality; VCF_DP: variant call format-read depth; VCF_GT: variant call format-genotype quality.



Data Preprocessing and Genotype Quality Filters

For all the variants, we have their unphased genotype information using the format of VCF_GT. To make the data processable for deep learning models, we encoded the VCF_GT data by creating categorical values to represent different types of genotype [27]. Specifically, 0 denotes that both allele values are reference alleles, 1 represents one allele value is a reference allele and the other one is the alternate allele, and 2 represents both are alternate alleles. For example, 0/0 is made as 0, 1/0 is made as 1, 1/1 is made as 2, and so on. Therefore, the variants consist of 3 categories: 0, 1, and 2. We used VCF_DP (read depth at a position for a sample) and VCF_CQ (conditional genotype quality) as a filter to control the genotype information quality (Figure 1A). We extracted the genotype information for each variant that has a read depth no less than 10 and genotype quality no less than 90 [31]. Therefore, the genotypes of read depth less than 10 (VCF_DP < 10) or genotype quality less than 90 (VCF_CQ < 90) were excluded. Since we only explored common variants in our study, we removed all the variants with occurrence frequency less than 1% over the whole data set, resulting in 153,347 variants selected as common variants after the genotype quality filters (Figure 1A). We used these common variants for our study. After selecting the common variants, the SSC samples were partitioned into 2 sets based on random sampling of individuals into a training set (80%) and a hold-out test set (20%). There was no overlap of individuals across the 2 partitions. The test set was only used after model fitting to assess performance.

Identifying Contributory Common Genetic Variants

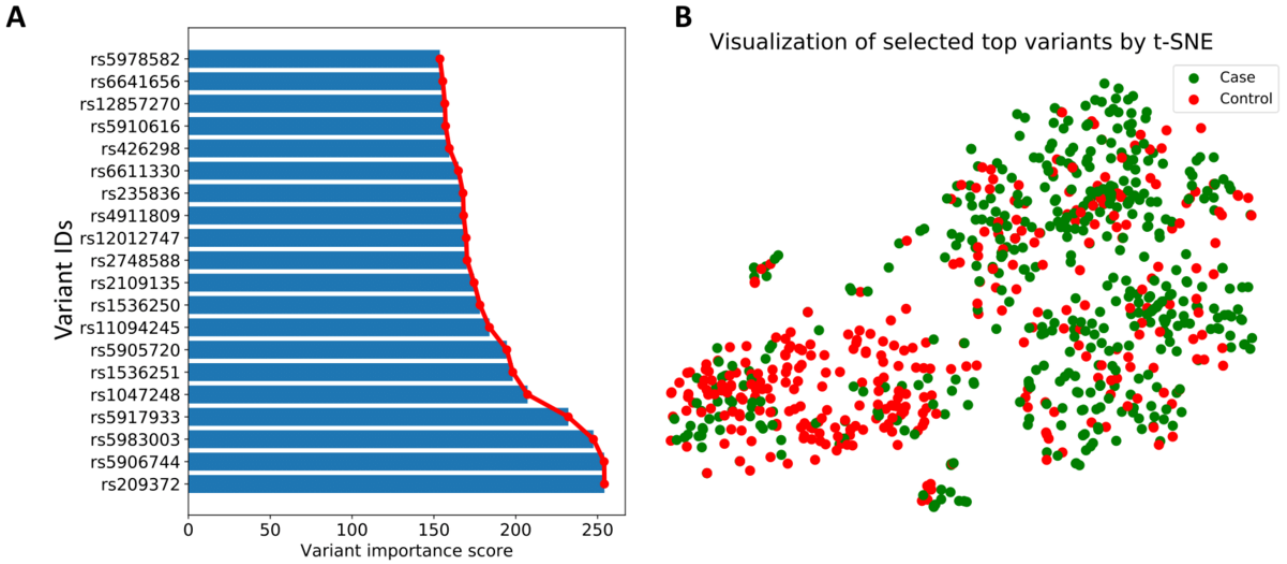
As the number of variants was too large to apply deep learning models directly, to construct the features for the deep learning models, we used feature selection to reduce variant dimension (Figure 1B). Feature selection is one of the core concepts in machine learning that hugely impacts the performance of a model [32-35]. The data features that are used to train machine learning models have a huge influence on the performance that we can achieve. Therefore, our hypothesis is that not all variables contribute to the predictive performance of the models we built. Variant selection is the process wherein we automatically select those features that contribute most to our prediction accuracy and are considered as contributory variants to ASD diagnosis. Therefore, significant common variation selection was applied because variant selection is the process of removing redundant or irrelevant features from the original

data set to reduce overfitting. To this end, we analyzed the importance scores of the common variants that are related with ASD development mechanisms in the training data set. For each individual, a 153,347-dimensional vector was constructed, corresponding to 153,347 common variants identified from the data preprocessing. Chi-square test was applied to evaluate the importance of each variant to distinguish the class in order to select the most significant common variants. Given the training data D, we estimated the following quantity for each variant and ranked them by their scores:

where N is the observed frequency in D and E is the expected frequency, e_t takes the value 1 if the training data contains term t and 0; otherwise, e_c takes the value 1 if the training data is in class c and 0 otherwise. For each variant, a corresponding high score indicates that the null hypothesis H_0 of independence (meaning the individual's category has no influence over the term's frequency) should be rejected and the occurrence of the variant and class are dependent. In this case, we select the variant for the ASD diagnosis prediction. We used the implementation from scikit-learn [36] for "Chi-Square Feature Selection" with default settings.

By calculating the chi-square scores for all the variants, we can rank the variants by the chi-square scores and then choose the top ranked variants as significant variants for model training. Figure 2A lists the variant importance of the high scoring 20 features (variants) that are selected via the chi-square test. In Figure 2A, while the Y-axis corresponds to variant IDs of the variants, the X-axis corresponds to relative importance, which is calculated using the chi-square score. We selected the top 100 most significant variants as inputs to train a deep learning classifier. Therefore, the number of input contributory variants to our classifier after selection was 100 for ASD prediction. In order to analyze whether the variation data can be divided into 2 clusters representing control and ASD cases, the first 2 groups of data were obtained using t-distributed stochastic neighbor embedding (t-SNE) as an unsupervised learning approach. The visualization of clusters for the top 100 variants using t-SNE in both case group and control group is shown in Figure 2B. From the visualization, accurate genetic classification of control group versus ASD is possible using 100 common variants determined to be highly significant. Therefore, for each individual in the training set, a 100-dimensional input vector was constructed corresponding to 100 selected significant common variants for training a deep learning model.

Figure 2. A. Variants with high relative importance scores in chi-square test. The Y-axis corresponds to variant IDs of these variants, and the X-axis corresponds to the relative importance values of the corresponding variants. B. Visualization of the top 100 selected significantly common variants using t-distributed stochastic neighbor embedding. Different colors represent different classes (ie, case and control). This visualization indicates that the 2 groups are differentiable using the selected top common variants. t-SNE: t-distributed stochastic neighbor embedding.



DeepAutism Architecture

The overall framework of the proposed DeepAutism (Figure 1) consists of 3 components, namely, data preprocessing, variant selection, and neural network classifier. Figure 1C illustrates the convolutional neural network (CNN) architecture. We used Keras and TensorFlow version 2.0 for constructing and training the CNN model. We used a block of two 1D convolutional layers, followed by a max-pooling layer to generate feature maps that contain only the most important features. The max-pooling layer is followed by a dropout layer to avoid overfitting the data. Then, the learned feature maps are combined using a fully connected layer. The final layer contains a sigmoid function to produce probabilities of output from 0 to 1, with the diseased group belonging to class 1 and the control set belonging to class 0. All the parameters, including the weights and biases of hidden layers, are learned through backpropagation [37]. The detailed network topology used in our CNN architecture is shown in Multimedia Appendix 1.

DeepAutism Training and Evaluation

For training, DeepAutism uses a set of selected common variants (top 100 significant common variants) to estimate the probabilities of an individual belonging to control case or

autism. For a set of variants *v* from a testing individual, DeepAutism computes a probability *p(v)* using 4 states:

$$p(v) = Sigmoid(netW(pool(ReLUb(convf(v)))))$$

The sigmoid function is used for computing probabilities of a set of variants v belonging to either control group or autism group, and the produced probabilities are from 0 to 1, with the control set belonging to class 0 and the ASD group belonging to class 1. The convolution stage (*convf*) scans a set of filters as feature maps across the variants. Each neuron consists of a rectified linear unit (ReLU) activation function to introduce nonlinearity between 2 neural networks. The pooling layer only picks the maximum values from the convolved feature maps. Since the variant data are categorical values, one-hot encoding is employed to ensure that the DeepAutism model is unbiased and does not favor one genotype over the others. The DeepAutism is then trained using mini-batch gradient descent by backpropagation algorithm [37]. The performance was evaluated using the area under the receiver operating characteristic curve (AUC). We also used the most common procedure for evaluating classifiers for ASD prediction, including accuracy, sensitivity (recall), specificity (precision), F1-score, and false discovery rate, in which the lower value indicates better performance to evaluate the classifiers for ASD diagnosis.

Baseline Methods to Compare the Effectiveness of DeepAutism

Apart from CNNs, we also employed conventional machine learning techniques to evaluate the effectiveness of DeepAutism for classifying autism diagnosis. The conventional machine learning models that we compared were random forests, logistic regression, and Naive Bayes. We used the same training and test data (with the selected 100 common variants) for the conventional machine learning models as used for the DeepAutism model, aiming to evaluate whether the CNN model outperforms other machine learning classifiers. To evaluate whether the selected top 100 common variants are significant for ASD diagnosis, we also compared the chi-square–based variant selection method with random variant selection by using the same training and test data sets. We randomly selected 100 common variants as inputs that were fed into both DeepAutism and conventional machine learning models to compare the changes in their performance.

Results

Identification of Contributory Variants and Genes

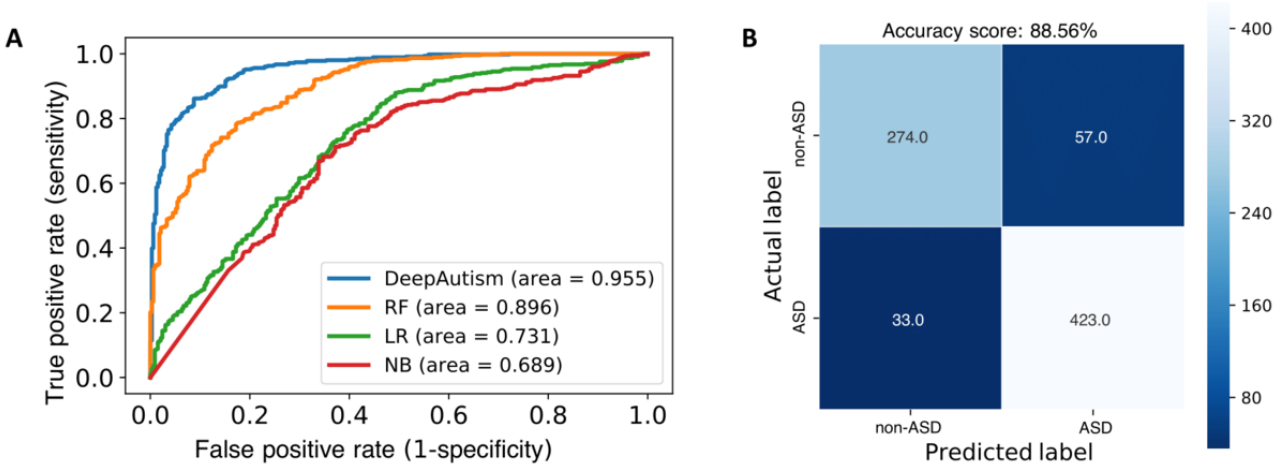
Statistical analyses focused on the selected top 100 common variants, which most significantly contributed to the classifiers of ASD. Of the 100 common variants within our classifier, 66% are exonic mutations and 23% are intronic mutations, while small proportions are splicing mutations or from an untranslated region. Within the 66% exonic mutations, about half are synonymous single nucleotide variants and about half are nonsynonymous single nucleotide variants. It is important to point out that the selected contributory common variants were significantly enriched on chromosome X while chromosome Y is also discriminatory in identifying individuals with ASD from individuals without ASD.

A number of variants were populated by the same genes. Related to the contributory common variants, the statistically significant genes were *ARSD*, *MAGEB16*, and *MXRA5*. There are 18 common variants in the *ARSD* gene. *ARSD* is a protein-coding gene and is located within a cluster of similar arylsulfatase genes on chromosome X, while a related pseudogene has been identified in the pseudoautosomal region of chromosome Y. Variants rs209372, rs2109135, and rs1047248 in 3 genes, namely, *NRK*, *TLR8*, and *MAGEA4*, respectively, have the highest scores in determining an individual's classification as with ASD or with no ASD.

Deep Learning Performance Based Upon Contributory Common Variants

After the training phase was over, we picked the same common variations from the test data for each individual. We used the rest of the 787 samples for testing. Based on the trained DeepAutism model, each test individual was predicted the probabilities of belonging to the control group or the diseased group. The deep learning model was extremely accurate in classification of the holdout test set with an AUC of 0.955 (Figure 3A). Figure 3B describes the performance of the DeepAutism classifier on the test data. DeepAutism predicted ASD in 423 samples out of 456 samples with ASD.

Figure 3. A. The area under the receiver operating characteristic curve of DeepAutism, random forest, logistic regression, and Naive Bayes for predicting autism spectrum disorder diagnosis based on the selected top 100 significantly common variants on the test data. B. The visualization table that describes the performance of the DeepAutism classifier on the test data. DeepAutism correctly predicted 697 out of 787 total samples and correctly predicted autism spectrum disorder in 423 samples out of 456 samples with autism spectrum disorders. AUC: area under the receiver operating characteristic curve; ASD: autism spectrum disorder; NB: Naive Bayes; LR: logistic regression; RF: random forest.



Apart from deep learning, we also employed Naive Bayes, logistic regression, support vector machine, random forest, and deep neural network classifiers to compare the prediction of ASD diagnosis. We applied five-fold cross-validation to evaluate the selected significant common variants. Our classifier performed better than the conventional machine learning techniques in terms of AUC, accuracy, specificity, sensitivity, and F1-score. As shown in Table 1, accuracy was 0.886 in the case of DeepAutism, followed by 0.808 for random forest in the same test data set for ASD diagnosis prediction. DeepAutism also yielded the best sensitivity of 0.881 for prediction of ASD and best specificity of 0.893 for non-ASD prediction. The false positive (discriminatory) rate is minimum for DeepAutism with 7% compared with other machine learning techniques. These results are shown in Table 1.

Table 1. Performance of the classifiers with respect to accuracy, sensitivity, specificity, F1-score, and false discovery rate on test sets.[a]

Model	Accuracy	Sensitivity	Specificity	F1-score	False discovery rate
DeepAutism	*0.886*	*0.881*	*0.893*	*0.905*	*0.072*
Naive Bayes	0.679	0.706	0.633	0.733	0.237
Random forest	0.808	0.785	0.857	0.848	0.079
Logistic regression	0.704	0.715	0.683	0.761	0.186
Support vector machine	0.789	0.773	0.821	0.831	0.101
Deep neural network	0.804	0.766	0.885	0.842	0.073

[a]Italicized data demonstrate the best performance; DeepAutism outperformed other models on all the metrics.

Performance Using Randomly Selected Common Variants for ASD Diagnosis

We assessed the classification performance by using randomly picked 100 common variants as inputs to train classifiers. We used the same training and test data as in the above experiment. As shown in Table 2, when the classifiers classify ASD using randomly selected common variants, all the classifiers achieved reduced performance compared to using selected significant common variants. For instance, the AUC and accuracy of DeepAutism dramatically dropped from 0.955 to 0.670 and from 0.885 to 0.689, respectively. The random 100 common variants yielded accuracy of 0.454 and 0.583 using Naive Bayes and logistic regression classifiers, respectively, which is like random guessing. This revealed that the random 100 common variants are not discriminative in distinguishing ASD diagnosis. These results suggest that variant selection is important for identifying significant common variants that are more correlated and significant in improving the classification accuracy.

Table 2. Performance of the classifiers with respect to area under receiver operating characteristic curve, accuracy, sensitivity, specificity, F1-score, and false discovery rate on test sets with randomly picked 100 common variants.[a]

Model	Area under receiver operating characteristic curve	Accuracy	Sensitivity	Specificity	F1-score	False discovery rate
DeepAutism	0.670	*0.689*	0.685	0.697	*0.755*	0.145
Naive Bayes	0.556	0.454	*0.717*	0.432	0.166	0.906
Random forest	*0.701*	0.629	0.612	*0.855*	0.754	*0.018*
Logistic regression	0.571	0.583	0.598	0.489	0.704	0.143
Support vector machine	0.672	0.679	0.633	0.571	0.696	0.139
Deep neural network	0.656	0.677	0.681	0.702	0.733	0.143

[a]Italicized data show the best performance; the performance of all models became worse on all the metrics with randomly selected common variants.

Discussion

Predicting ASD based on genetic data is challenging. Using common variant analysis, we generated a genetic diagnostic classifier (DeepAutism) based on a deep learning architecture using 100 significant common variants, and we accurately distinguished ASD from controls within the SSC data set. The diagnostic classifier was able to correctly classify individuals with ASD with an accuracy of 88.6% and an AUC of 0.955. Our findings showed that the sensitivity and specificity of the classifier when applied to identify ASD were 88% and 89%, respectively. It is notable that the sensitivity for identifying cases is highly desirable for screening purposes. We also investigated the classification performance of different approaches and the corresponding proportion of subjects who did not have ASD who could be reliably classified as controls. DeepAutism can be suggested as an alternative to conventional shallow machine learning approaches. In the comparisons among the classifiers, DeepAutism performed the best, followed by random forest. Both these classifiers are nonlinear models. Therefore, the causes of ASD are not a simple linear combination of common variants.

Interestingly, when we altered the classifier by using randomly selected 100 common variants, the AUC and accuracy of DeepAutism reduced to 0.670 and 0.689, respectively. The performance became worse because irrelevant variants can include noisy data, thereby affecting the classification accuracy negatively. This verifies the significance of selecting common variants and greatly adds strength to our original findings. Our results suggest that common variants may contribute to ASD diagnosis. A study [18] has shown that the genetic architecture of ASD is contributed by inherited common variants, which supports our findings. The common variants contributing most to the diagnosis in our classifier corresponded to genes on chromosome X. This suggests that ASD is associated with gender. As ASD is strongly biased toward males with ratios of 4:1 (male:female) [38] and statistics have also shown that ASD has a higher prevalence in males than in females [39], mutations in the genes on the X chromosome may explain the increased prevalence of autism in boys compared to that in girls. Thus, this supports our finding that gender bias affects individuals with autism.

In our findings, *ARSD*, *MAGEB16*, and *MXRA5* genes were found to have a high contributory effect on ASD. *ARSD* is located within a cluster of similar arylsulfatase genes on chromosome X. *ARSD* is clinically heterogeneous and is likely to result from mutations in developmental genes or from regulating transcription factors [40]. *ARSD* has already been reported to be related to ASD or Asperger's syndrome [41]. The cytogenetic location of *ARSD* is Xp22.33, and this location significantly contributes to ASD, as shown in the SFARI Gene Database [42]. These regions play a role in neurodevelopment disorders [31,43-48]. Although we used common variants as features in our classifier, we also found that prevalence of X-chromosome copy number variations contribute to ASD. *MAGEB16* is also a protein-coding gene, which is located on Xp21.1. *MAGEB16* has been implicated in syndromic X-linked intellectual disability and neurodevelopmental disorders. It has also been reported to be associated with autistic disorders [49]. *MXRA5* is a protein-coding gene and encodes a protein that forms the extracellular matrix structural constituent. It is involved in the response to transforming growth factor beta and has a pseudogene on chromosome Y. An association has been curated linking *MXRA5* and an autistic disorder in *Pan paniscus*. Although mutations in *SHANK3* have been identified in multiple individuals with ASD, most of the mutations are rare variants and not common variants, where the ratio between rare variants and common variants is 230:9 according to the SFARI Gene Database [50].

ASD is a complex behavioral disorder with a strong genetic influence [51]. Diagnosing ASD can be difficult because there is no medical test (such as a blood test) to diagnose this disorder. Although the majority of studies toward biomarker identification for autism have focused on rare genetic variants, we have proven that common genetic variants are also informative with respect to the identification of ASD. In our study, our genetic classifier obtained a high level of diagnostic accuracy, thereby demonstrating that genetic biomarkers can correctly identify individuals with ASD from individuals without ASD. Common variants can play a very important role in screening ASD at an early stage. We identified a few genes with various common variants that could determine whether an individual fell within the case or control group. Our results demonstrate the value of a data-driven approach for the identification of significant common variants and a deep learning method for ASD diagnosis. Overall, these findings indicate that a common

variant–based test may allow for early identification of ASD. A genetic predictive classifier as described here may be a tool for ASD screening at birth to provide probability estimates of ASD.

Although our approach for identifying autism based on the selected common variants achieves high accuracy, some limitations exist that need improvement in the future work: (1) the experiments were conducted on the SSC dataset; however, more datasets could be used to evaluate the proposed method and the selected common variants and (2) the proposed algorithm, based on CNN, is a straightforward solution for identifying autism from nonautism; however, more state-of-the-art classifiers could be applied to this ASD classification problem.

While the proposed DeepAutism approach has achieved great success in ASD identification with promising empirical results, we would still like to explore several important directions on DeepAutism in the future. First, we plan to further design an advanced deep learning algorithm that can handle high-dimensional features and output the feature importance for variant selection. By using the designed model, we can select significant variants and classify autistic individuals simultaneously as an end-to-end framework. Second, we will evaluate the proposed method on 2 more distinct ASD cohorts: (1) Simons Foundation Powering Autism Research for Knowledge data and (2) Autism Speaks MMSSNG cohort. We will also validate our algorithms with the UK Biobank clinical and genomic data. Third, we will investigate the full sequences of coding and noncoding regions of the genome between probands and unaffected siblings to explore all of the components in the genetic architecture of ASD.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Architecture of DeepAutism.
[DOCX File , 106 KB - medinform_v9i4e24754_app1.docx]

References

1. Rylaarsdam L, Guemez-Gamboa A. Genetic Causes and Modifiers of Autism Spectrum Disorder. Front Cell Neurosci 2019;13:385 [FREE Full text] [doi: 10.3389/fncel.2019.00385] [Medline: 31481879]
2. Frazier TW, Thompson L, Youngstrom EA, Law P, Hardan AY, Eng C, et al. A twin study of heritable and shared environmental contributions to autism. J Autism Dev Disord 2014 Aug;44(8):2013-2025 [FREE Full text] [doi: 10.1007/s10803-014-2081-2] [Medline: 24604525]
3. Sutton H. Autism caused mostly by genetics, according to study. Disability Compliance for Higher Education 2019 Aug 22;25(2):9-9 [FREE Full text] [doi: 10.1002/dhe.30707]
4. McDonald NM, Senturk D, Scheffler A, Brian JA, Carver LJ, Charman T, et al. Developmental Trajectories of Infants With Multiplex Family Risk for Autism: A Baby Siblings Research Consortium Study. JAMA Neurol 2020 Jan 01;77(1):73-81 [FREE Full text] [doi: 10.1001/jamaneurol.2019.3341] [Medline: 31589284]
5. de la Torre-Ubieta L, Won H, Stein JL, Geschwind DH. Advancing the understanding of autism disease mechanisms through genetics. Nat Med 2016 Apr;22(4):345-361 [FREE Full text] [doi: 10.1038/nm.4071] [Medline: 27050589]
6. Derecki NC, Cronk JC, Lu Z, Xu E, Abbott SBG, Guyenet PG, et al. Wild-type microglia arrest pathology in a mouse model of Rett syndrome. Nature 2012 Mar 18;484(7392):105-109 [FREE Full text] [doi: 10.1038/nature10907] [Medline: 22425995]
7. Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. Nature 2020 Oct;586(7827):80-86. [doi: 10.1038/s41586-020-2579-z] [Medline: 32717741]
8. Maestrini E, Pagnamenta AT, Lamb JA, Bacchelli E, Sykes NH, Sousa I, IMGSAC. High-density SNP association study and copy number variation analysis of the AUTS1 and AUTS5 loci implicate the IMMP2L-DOCK4 gene region in autism susceptibility. Mol Psychiatry 2010 Sep;15(9):954-968 [FREE Full text] [doi: 10.1038/mp.2009.34] [Medline: 19401682]
9. Bai D, Yip BHK, Windham GC, Sourander A, Francis R, Yoffe R, et al. Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort. JAMA Psychiatry 2019 Oct 01;76(10):1035-1043 [FREE Full text] [doi: 10.1001/jamapsychiatry.2019.1411] [Medline: 31314057]
10. Lintas C, Picinelli C, Piras IS, Sacco R, Brogna C, Persico AM. Copy number variation in 19 Italian multiplex families with autism spectrum disorder: Importance of synaptic and neurite elongation genes. Am J Med Genet B Neuropsychiatr Genet 2017 Jul;174(5):547-556. [doi: 10.1002/ajmg.b.32537] [Medline: 28304131]
11. Sahin NT, Keshav NU, Salisbury JP, Vahabzadeh A. Second Version of Google Glass as a Wearable Socio-Affective Aid: Positive School Desirability, High Usability, and Theoretical Framework in a Sample of Children with Autism. JMIR Hum Factors 2018 Jan 04;5(1):e1 [FREE Full text] [doi: 10.2196/humanfactors.8785] [Medline: 29301738]

12. Ahmed KL, Simon AR, Dempsey JR, Samaco RC, Goin-Kochel RP. Evaluating Two Common Strategies for Research Participant Recruitment Into Autism Studies: Observational Study. J Med Internet Res 2020 Sep 24;22(9):e16752 [FREE Full text] [doi: 10.2196/16752] [Medline: 32969826]
13. Siu M, Butcher DT, Turinsky AL, Cytrynbaum C, Stavropoulos DJ, Walker S, et al. Functional DNA methylation signatures for autism spectrum disorder genomic risk loci: 16p11.2 deletions and CHD8 variants. Clin Epigenetics 2019 Jul 16;11(1):103 [FREE Full text] [doi: 10.1186/s13148-019-0684-3] [Medline: 31311581]
14. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell 2017 Oct 19;171(3):710-722.e12 [FREE Full text] [doi: 10.1016/j.cell.2017.08.047] [Medline: 28965761]
15. Vorstman JAS, Parr JR, Moreno-De-Luca D, Anney RJL, Nurnberger JI, Hallmayer JF. Autism genetics: opportunities and challenges for clinical translation. Nat Rev Genet 2017 Jun;18(6):362-376. [doi: 10.1038/nrg.2017.4] [Medline: 28260791]
16. Ronemus M, Iossifov I, Levy D, Wigler M. The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet 2014 Feb;15(2):133-141. [doi: 10.1038/nrg3585] [Medline: 24430941]
17. Sanders S, He X, Willsey A, Ercan-Sencicek A, Samocha K, Cicek A, Autism Sequencing Consortium, et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. Neuron 2015 Sep 23;87(6):1215-1233 [FREE Full text] [doi: 10.1016/j.neuron.2015.09.016] [Medline: 26402605]
18. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nat Genet 2014 Aug;46(8):881-885 [FREE Full text] [doi: 10.1038/ng.3039] [Medline: 25038753]
19. Grove, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Autism Spectrum Disorder Working Group of the Psychiatric Genomics Consortium, BUPGEN, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, 23andMe Research Team, et al. Identification of common genetic risk variants for autism spectrum disorder. Nat Genet 2019 Mar;51(3):431-444 [FREE Full text] [doi: 10.1038/s41588-019-0344-8] [Medline: 30804558]
20. Thapar A, Rutter M. Genetic Advances in Autism. J Autism Dev Disord 2020 Sep 17. [doi: 10.1007/s10803-020-04685-z] [Medline: 32940822]
21. Chen JA, Peñagarikano O, Belgard TG, Swarup V, Geschwind DH. The emerging picture of autism spectrum disorder: genetics and pathology. Annu Rev Pathol 2015;10:111-144. [doi: 10.1146/annurev-pathol-012414-040405] [Medline: 25621659]
22. Skafidas E, Testa R, Zantomio D, Chana G, Everall IP, Pantelis C. Predicting the diagnosis of autism spectrum disorder using gene pathway analysis. Mol Psychiatry 2014 Apr;19(4):504-510 [FREE Full text] [doi: 10.1038/mp.2012.126] [Medline: 22965006]
23. Howsmon DP, Kruger U, Melnyk S, James SJ, Hahn J. Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. PLoS Comput Biol 2017 Mar;13(3):e1005385 [FREE Full text] [doi: 10.1371/journal.pcbi.1005385] [Medline: 28301476]
24. Amoedo A, Martnez-Costa MDP, Moreno E. An analysis of the communication strategies of Spanish commercial music networks on the web: http://los40.com, http://los40principales.com, http://cadena100.es, http://europafm.es and http://kissfm.es. Radio Journal: International Studies in Broadcast & Audio Media 2009 Feb 01;6(1):5-20 [FREE Full text] [doi: 10.1386/rajo.6.1.5_4]
25. Krishnan A, Zhang R, Yao V, Theesfeld CL, Wong AK, Tadych A, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat Neurosci 2016 Nov;19(11):1454-1462 [FREE Full text] [doi: 10.1038/nn.4353] [Medline: 27479844]
26. Chen T, Chen Y, Yuan M, Gerstein M, Li T, Liang H, et al. The Development of a Practical Artificial Intelligence Tool for Diagnosing and Evaluating Autism Spectrum Disorder: Multicenter Study. JMIR Med Inform 2020 May 08;8(5):e15767 [FREE Full text] [doi: 10.2196/15767] [Medline: 32041690]
27. Sundaram L, Bhat RR, Viswanath V, Li X. DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning. Hum Mutat 2017 Sep;38(9):1217-1224 [FREE Full text] [doi: 10.1002/humu.23272] [Medline: 28600868]
28. Moon SJ, Hwang J, Kana R, Torous J, Kim JW. Accuracy of Machine Learning Algorithms for the Diagnosis of Autism Spectrum Disorder: Systematic Review and Meta-Analysis of Brain Magnetic Resonance Imaging Studies. JMIR Ment Health 2019 Dec 20;6(12):e14108 [FREE Full text] [doi: 10.2196/14108] [Medline: 31562756]
29. Ben-Sasson A, Robins DL, Yom-Tov E. Risk Assessment for Parents Who Suspect Their Child Has Autism Spectrum Disorder: Machine Learning Approach. J Med Internet Res 2018 Apr 24;20(4):e134 [FREE Full text] [doi: 10.2196/jmir.9496] [Medline: 29691210]
30. Sullivan MO, Gallagher L, Heron EA. Gaining Insights into Aggressive Behaviour in Autism Spectrum Disorder Using Latent Profile Analysis. J Autism Dev Disord 2019 Oct;49(10):4209-4218 [FREE Full text] [doi: 10.1007/s10803-019-04129-3] [Medline: 31292900]
31. Doan RN, Lim ET, De Rubeis S, Betancur C, Cutler DJ, Chiocchetti AG, Autism Sequencing Consortium, et al. Recessive gene disruptions in autism spectrum disorder. Nat Genet 2019 Jul;51(7):1092-1098 [FREE Full text] [doi: 10.1038/s41588-019-0433-8] [Medline: 31209396]

32. Velusamy D, Ramasamy K. Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset. Comput Methods Programs Biomed 2021 Jan;198:105770. [doi: 10.1016/j.cmpb.2020.105770] [Medline: 33027698]
33. Jiang J, Cameron A, Yang M. Analysis of Massive Online Medical Consultation Service Data to Understand Physicians' Economic Return: Observational Data Mining Study. JMIR Med Inform 2020 Feb 18;8(2):e16765 [FREE Full text] [doi: 10.2196/16765] [Medline: 32069213]
34. Chikersal P, Doryab A, Tumminia M, Villalba DK, Dutcher JM, Liu X, et al. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing. ACM Trans Comput Hum Interact 2021 Feb;28(1):1-41. [doi: 10.1145/3422821]
35. Zhang Y, Zhou Y, Zhang D, Song W. A Stroke Risk Detection: Improving Hybrid Feature Selection Method. J Med Internet Res 2019 Apr 02;21(4):e12437 [FREE Full text] [doi: 10.2196/12437] [Medline: 30938684]
36. Scikit-learn: Machine learning in Python. URL: http://scikit-learn.org [accessed 2021-03-22]
37. Obeid, Dahne J, Christensen S, Howard S, Crawford T, Frey LJ, et al. Identifying and Predicting Intentional Self-Harm in Electronic Health Record Clinical Notes: Deep Learning Approach. JMIR Med Inform 2020 Jul 30;8(7):e17784 [FREE Full text] [doi: 10.2196/17784] [Medline: 32729840]
38. Cahill L. A New Link Between Autism and Masculinity. JAMA Psychiatry 2017 Apr 01;74(4):318. [doi: 10.1001/jamapsychiatry.2016.4066] [Medline: 28196210]
39. Hull L, Petrides KV, Mandy W. The Female Autism Phenotype and Camouflaging: a Narrative Review. Rev J Autism Dev Disord 2020 Jan 29;7(4):306-317. [doi: 10.1007/s40489-020-00197-9]
40. Turnpenny PD, Bulman MP, Frayling TM, Abu-Nasra TK, Garrett C, Hattersley AT, et al. A gene for autosomal recessive spondylocostal dysostosis maps to 19q13.1-q13.3. Am J Hum Genet 1999 Jul;65(1):175-182 [FREE Full text] [doi: 10.1086/302464] [Medline: 10364530]
41. ARSD: arylsulfatase D. URL: https://www.wikigenes.org/e/gene/e/414.html [accessed 2021-03-22]
42. Copy number variants/Xp22.33. SFARI Gene. URL: https://gene-archive.sfari.org/database/cnv/Xp22.33 [accessed 2021-03-22]
43. Willemsen MH, de Leeuw N, de Brouwer AP, Pfundt R, Hehir-Kwa JY, Yntema HG, et al. Interpretation of clinical relevance of X-chromosome copy number variations identified in a large cohort of individuals with cognitive disorders and/or congenital anomalies. Eur J Med Genet 2012 Nov;55(11):586-598. [doi: 10.1016/j.ejmg.2012.05.001] [Medline: 22796527]
44. Asadollahi R, Oneda B, Joset P, Azzarello-Burri S, Bartholdi D, Steindl K, et al. The clinical significance of small copy number variants in neurodevelopmental disorders. J Med Genet 2014 Oct;51(10):677-688 [FREE Full text] [doi: 10.1136/jmedgenet-2014-102588] [Medline: 25106414]
45. Kushima, Aleksic B, Nakatochi M, Shimamura T, Okada T, Uno Y, et al. Comparative Analyses of Copy-Number Variation in Autism Spectrum Disorder and Schizophrenia Reveal Etiological Overlap and Biological Insights. Cell Rep 2018 Sep 11;24(11):2838-2856 [FREE Full text] [doi: 10.1016/j.celrep.2018.08.022] [Medline: 30208311]
46. Rosenfeld JA, Ballif BC, Torchia BS, Sahoo T, Ravnan JB, Schultz R, et al. Copy number variations associated with autism spectrum disorders contribute to a spectrum of neurodevelopmental disorders. Genet Med 2010 Aug 30;12(11):694-702. [doi: 10.1097/gim.0b013e3181f0c5f3]
47. Edens A, Lyons M, Duron R, Dupont BR, Holden KR. Autism in two females with duplications involving Xp11.22-p11.23. Dev Med Child Neurol 2011 May;53(5):463-466 [FREE Full text] [doi: 10.1111/j.1469-8749.2010.03909.x] [Medline: 21418194]
48. Ben-David E, Granot-Hershkovitz E, Monderer-Rothkoff G, Lerer E, Levi S, Yaari M, et al. Identification of a functional rare variant in autism using genome-wide screen for monoallelic expression. Hum Mol Genet 2011 Sep 15;20(18):3632-3641. [doi: 10.1093/hmg/ddr283] [Medline: 21680558]
49. MAGEB16. Alliance of genome resources. URL: https://www.alliancegenome.org/gene/HGNC:21188 [accessed 2021-03-22]
50. SHANK3: SH3 and multiple ankyrin repeat domains 3. SFARI Gene. URL: https://gene.sfari.org/database/human-gene/SHANK3 [accessed 2021-03-22]
51. Yoo H. Genetics of Autism Spectrum Disorder: Current Status and Possible Clinical Applications. Exp Neurobiol 2015 Dec;24(4):257-272 [FREE Full text] [doi: 10.5607/en.2015.24.4.257] [Medline: 26713075]

Abbreviations

ASD: autism spectrum disorder
AUC: area under the receiver operating characteristic curve
CNN: convolutional neural network
SSC: Simons Simplex Collection
t-SNE: t-distributed stochastic neighbor embedding
VCF: variant call format
VCF_CQ: variant call format-conditional genotype quality

VCF_DP: variant call format-read depth
VCF_GT: variant call format-genotype quality

Edited by G Eysenbach; submitted 03.10.20; peer-reviewed by S Pang, F Li, M Manzanares; comments to author 23.11.20; revised version received 18.02.21; accepted 14.03.21; published 07.04.21.

Please cite as:
Wang H, Avillach P
Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning
JMIR Med Inform 2021;9(4):e24754
URL: https://medinform.jmir.org/2021/4/e24754
doi:10.2196/24754
PMID:33714937

Original Paper

Using General-purpose Sentiment Lexicons for Suicide Risk Assessment in Electronic Health Records: Corpus-Based Analysis

André Bittar[1], PhD; Sumithra Velupillai[1], PhD; Angus Roberts[1], PhD; Rina Dutta[1,2], PhD

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

[2]South London and Maudsley NHS Foundation Trust, London, United Kingdom

Corresponding Author:
André Bittar, PhD
Institute of Psychiatry, Psychology and Neuroscience
King's College London
16 De Crespigny Park
London, SE5 8AF
United Kingdom
Phone: 44 (0)20 3228 8553
Email: andre.bittar@kcl.ac.uk

Abstract

Background: Suicide is a serious public health issue, accounting for 1.4% of all deaths worldwide. Current risk assessment tools are reported as performing little better than chance in predicting suicide. New methods for studying dynamic features in electronic health records (EHRs) are being increasingly explored. One avenue of research involves using sentiment analysis to examine clinicians' subjective judgments when reporting on patients. Several recent studies have used general-purpose sentiment analysis tools to automatically identify negative and positive words within EHRs to test correlations between sentiment extracted from the texts and specific medical outcomes (eg, risk of suicide or in-hospital mortality). However, little attention has been paid to analyzing the specific words identified by general-purpose sentiment lexicons when applied to EHR corpora.

Objective: This study aims to quantitatively and qualitatively evaluate the coverage of six general-purpose sentiment lexicons against a corpus of EHR texts to ascertain the extent to which such lexical resources are fit for use in suicide risk assessment.

Methods: The data for this study were a corpus of 198,451 EHR texts made up of two subcorpora drawn from a 1:4 case-control study comparing clinical notes written over the period leading up to a suicide attempt (cases, n=2913) with those not preceding such an attempt (controls, n=14,727). We calculated word frequency distributions within each subcorpus to identify representative keywords for both the case and control subcorpora. We quantified the relative coverage of the 6 lexicons with respect to this list of representative keywords in terms of weighted precision, recall, and F score.

Results: The six lexicons achieved reasonable precision (0.53-0.68) but very low recall (0.04-0.36). Many of the most representative keywords in the suicide-related (case) subcorpus were not identified by any of the lexicons. The sentiment-bearing status of these keywords for this use case is thus doubtful.

Conclusions: Our findings indicate that these 6 sentiment lexicons are not optimal for use in suicide risk assessment. We propose a set of guidelines for the creation of more suitable lexical resources for distinguishing suicide-related from non–suicide-related EHR texts.

(JMIR Med Inform 2021;9(4):e22397) doi:10.2196/22397

KEYWORDS

psychiatry; suicide; suicide, attempted; risk assessment; electronic health records; sentiment analysis; natural language processing; corpus linguistics

Introduction

Background

The World Health Organization reports that suicide accounts for 1.4% of all deaths globally and is the 18th leading cause of death worldwide [1]. Prior history of suicide attempts is the most robust risk factor for completed suicide, and those requiring hospitalization are at the most serious end of the spectrum [2]. However, current methods for assessing a patient's risk of attempting suicide are reported to perform little better

than chance [3]. Therefore, new methods to understand dynamic features from electronic health records (EHRs) before a hospitalized suicide attempt, distinguishing such periods from clinical narratives at other times, would be of potential clinical utility [4].

EHRs contain structured patient data (eg, age, sex, and ethnicity) and unstructured text that make up the clinical narrative (eg, out-patient letters, event notes from meetings and phone calls with patients or carers, and discharge summaries). Unstructured text is of particular importance in mental health, as much of what is recorded about patients follows face-to-face assessments by clinicians, whose observations and judgments about a patient's experiences and presentation are inevitably influenced by their own training, experience, and implicit biases, and these judgments have a degree of subjectivity when they record this in the clinical narrative [5].

The automatic identification and analysis of subjective judgments in text is known as sentiment analysis [6,7]. This process typically involves the classification of words as expressing either positive or negative polarity, and numerous resources have been developed for this task in nonclinical domains, such as customer reviews [8-11] and social media [12-14]. Research efforts have also focused on the analysis of sentiment within health care–related texts, such as patient feedback forms [15,16], online forums [17], and social networks [18,19].

Recent work has sought to assess the utility of sentiment lexicons for the analysis of subjective judgments in clinical narratives. McCoy et al [20] used a general-domain sentiment analysis tool to extract word polarity features to model the risk of readmission and mortality. The same tool was later used to examine the correlation between word polarity and the risk of suicide attempts [21]. Most recently, Weissman et al [22] carried out a thorough evaluation of six general-domain sentiment analysis tools in predicting the risk of in-hospital mortality of patients in intensive care, tracking the progression of sentiment in clinical notes over time. They concluded that general-domain sentiment tools are not suited to the processing of clinical texts and that domain-specific resources need to be developed. Work in this direction is beginning to emerge [23-25].

These studies have mostly focused on testing the correlation between automatically extracted sentiment values and specific clinical outcomes. However, to our knowledge, there has been no close examination of the terms mapped by general-domain sentiment analysis tools when applied to clinical texts.

Objectives

Focusing on words with negative and positive polarity, we aimed to determine the coverage of 6 general-purpose sentiment lexicons when applied to a corpus of EHR texts of 2 groups of patients seen by mental health services: (1) patients who had attempted suicide and were hospitalized (cases) and (2) patients with no history of attempted suicide (controls). Adopting methods used in corpus linguistics, we first sought to identify the words that are most representative of the clinical narratives of cases and controls. We then aimed to test the coverage of each sentiment lexicon by comparing these 2 sets of representative words. We sought to ascertain the extent to which these 2 sets of representative words contained general-purpose sentiment words and to what extent these 2 sets contained additional sentiment words not included in the general-purpose lexicons.

Methods

Corpus Analysis

Clinical Cohort

We studied deidentified EHRs of over 250,000 patients from the South London and Maudsley National Health Service Foundation Trust using the Clinical Record Interactive Search (CRIS) database, comprising over 3.5 million text documents [26]. CRIS has been linked with national hospital admission data within a secure *safe haven*, allowing hospital admission information to be extracted. The deidentified CRIS database has received ethical approval for secondary analysis: Oxford REC C, reference 18/SC/0372. Access is granted upon request to authorized researchers working on projects that have received prior approval from the CRIS Oversight Committee. The data presented in this study can be viewed within the secure system firewall.

Our data set was derived from the EHRs of 17,640 patients. It consisted of 4235 suicide attempt–related (case) admissions and 16,940 nonsuicide attempt–related (control) admissions, sampled according to a 1:4 case-control ratio. Cases were defined as any admission (acute physical or specialist mental health) where there was a suicide attempt (indicated by any of the following codes from the International Classification of Diseases (ICD-10): X6*, X7*, X80-4*, Y1*, Y2*, Y30-4*, and Y87*) with the admission lasting at least 24 hours. Admissions starting on or after April 1, 2006, and ending before or including March 31, 2017, were considered. Case admissions that had at least one document in the 30 days up to and including the date of the suicide attempt were retained. We also removed admissions with empty documents (text from scanned documents is not always available in CRIS), resulting in a total of 4235 suicide-related admissions. Controls did not have any of the specified ICD-10 codes in the given period, were matched by sex, had to be alive at the admission start date of the corresponding case, and were matched to the same age group (5-year age bands: <16, 16-19, 20-24 to 80-84, and >85 years). Each control also had at least one document in the 30 days up to and including the date of the suicide attempt of the matched case. The controls were chosen to be representative (in terms of age and sex) of the population from which the cases were drawn, and the ratio was based on the epidemiological principle that little statistical power is gained by further increasing the number of controls beyond approximately 4 per case [27]. The key descriptive characteristics of the cohort are presented in Table 1.

Table 1. Cohort patient- and admission-level statistics.

Unit of observation	Cases	Controls
Patients, n (%)	2913 (16.51)	14,727 (83.49)
Female	1730 (59.39)	8971 (60.92)
Male	1183 (40.61)	5756 (39.08)
Admissions, n (%)	4235 (20.00)	16,940 (80.00)
Female	2598 (61.35)	10,392 (61.35)
Male	1637 (38.65)	6548 (38.65)
Age (years), mean (SD)	34.4 (15.3)	34.4 (15.4)

EHR Corpus

Our corpus comprised all EHR texts for each of the 2 subgroups in our clinical cohort: (1) suicidal case admissions and (2) nonsuicidal controls.

Our use of a 1:4 case-control study design for admissions means we expect a disparity in document number and word count between subcorpora. However, there are only 77.92% (55,643/71,404) more control documents (n=127,047) than case documents (n=71,404), rather than the 300% difference that might be expected for 1:4 sampling of random patients. Following data preprocessing (refer to the *Data Preparation* subsection), the mean lexical word count for case documents (n=117.4) is higher than that for control documents (n=103.9), so that the overall word (token) count ratio is not 1:4 but approximately 1:1.6, whereas the mean unique word (type) count ratio is approximately 1.5. The basic descriptive statistics for the corpus are shown in Table 2. The distribution of documents per patient followed a non-normal distribution, as shown in Multimedia Appendix 1.

Table 2. Electronic health record corpus descriptive statistics.

Unit of observation	Cases	Controls	Total
Word tokens, n	8,385,643	13,198,250	21,583,893
Word types, n	109,024	162,696	206,866
Type-token ratio[a], %	1.30	1.23	0.96
Documents, n	71,404	127,047	198,451
Number of words per document, mean (SD)	117.4 (219.1)	103.9 (252.7)	108.8 (241.3)

[a]Type-token ratio = number of word types / number of word tokens × 100.

Data Preparation

All texts were preprocessed using the Natural Language Processing (NLP) library spaCy (v2.0.12) [28], applying the following steps: word tokenization, part-of-speech tagging, and lemmatization (to use the base form of words). We removed stop words using the Natural Language ToolKit [29] stop words list for English and lowercased all words for our analyses. All codes were made available on GitHub [30].

Identifying Representative Keywords

To answer our questions concerning the coverage of each lexicon, we adopted methods based on word frequency distributions, commonly used in corpus linguistics, as described further in Multimedia Appendix 1 (C) [31-34]. We first determined which *keywords* were most *representative* of each subcorpus (suicidal case admission texts and nonsuicidal control texts) by calculating the relative word frequency ratios between subcorpora. Following recommendations from previous research in corpus linguistics [31-33] and given the non-normal distribution of documents between patients, we then applied the nonparametric Mann-Whitney *U* test to determine the statistical significance of word frequency differences (*FreqDiff (w)* for a given word w) between subcorpora. We only retained words that occurred in both the case and control subcorpora, leaving a total of 64,854 unique token types. Words appearing in only one or other subcorpora were relatively infrequent compared with those that were common to both subcorpora. For example, the most frequent case-only keywords were identifying initials, with a maximum frequency of 20.2 words per million (wpm), whereas the most frequent control-only keywords were persons' names, with a maximum frequency of 34.4 wpm.

Sentiment Lexicon Analysis

Sentiment Lexicons

We examined six different sentiment lexicons that were developed for nonclinical domains. Various dimensions of sentiment and affect have been studied, including emotion, valence-arousal-dominance, and polarity. We focused solely on lexicons that represent this last aspect, that is, negative and positive sentiment polarity. Along with assigning negative and positive polarity, some sentiment analysis tools also assign a value for words that do not convey semantic polarity (ie, *neutral* words). However, we only considered words that express positive and negative sentiments, as not all the lexicons in this study contain neutral terms. Therefore, we filtered out any

neutral words. Furthermore, for the sake of comparison, we only examined binary sentiment values rather than degree scores, which only some lexicons provide. We selected the following lexicons for this study: AFINN [35], the NRC Emotion Lexicon (commonly known as EmoLex) [36], Linguistic Inquiry and Word Count (LIWC) [37], the Opinion lexicon [9], the Pattern lexicon [38], and SentiWordNet [39]. The lexicons differ in terms of the forms they contain (words, lemmas, and regular expressions). We applied each one *as-is* to the appropriately preprocessed corpus (eg, words or lemmas) to compare them, as they have been used in other studies. We provide details of the lexicons, including preprocessing and filtering, in Multimedia Appendix 1 (B) [9,35-44]. Table 3 summarizes some of the main characteristics of each of these lexicons, including size before (original size) and after (filtered size) filtering out neutral entries.

Table 3. Characteristics of the 6 sentiment lexicons.

Lexicon	Source	Automatic term selection	Intended domain	Term type	Original size (entries), n	Filtered size (number of entries), n (%)
AFINN	Various web-based word lists	No	Microblogs	Word forms	3478	3478 (100.00)
EmoLex	Macquarie Thesaurus, General Inquirer, WordNet	No	General	Word forms	14,182	5555 (39.17)
LIWC[a]	Various dictionaries and thesauruses	No	Personal narratives	Word forms and regular expressions	1371	1371 (100.00)
Opinion	Web crawl of product reviews	Yes	Product reviews	Word forms	6789	6789 (100.00)
Pattern	Subset of WordNet	No	Product reviews	Lemmas+POS[b]	2896	2293 (79.18)
SentiWordNet	WordNet	Yes	General	Synset Lemmas+POS	117,659	39,746 (33.78)

[a]LIWC: Linguistic Inquiry and Word Count.

[b]POS: part of speech.

Lexicon Coverage

We assessed the coverage of each lexicon in three different ways:

1. *Global coverage*: The percentage of sentiment-bearing lexical entries that appeared in the list of (unique) words for each subcorpus. Further details are provided in Multimedia Appendix 1 (D).
2. *Keyword coverage*: The proportion of case and control keywords covered by the sentiment-bearing terms of a lexicon. First, we calculated the percentage of keywords identified by each lexicon for each subcorpus. Second, we used metrics common to information retrieval, namely, weighted precision (P_w), recall (R_w), and *F* score (F_w), which we calculated for each lexicon across the unordered set of all keywords, using word ranking as the weighting. Details of our calculations, including formulae, are provided in Multimedia Appendix 1 (D). A lexicon's precision shows how many case keywords it correctly identifies as a proportion of all the keywords it contains. The inclusion of control keywords in a lexicon, therefore, penalizes precision. In contrast, recall indicates the number of case keywords that the lexicon correctly identifies from the entire list of case keywords. The absence of case keywords from a lexicon results in a penalty on recall. *F*score provides a combination of the preceding 2 metrics and an overall quantified evaluation of a lexicon's keyword coverage.
3. *Sentiment coverage*: The sentiment polarity (positive or negative) that lexicons assigned to matched keywords for each subcorpus.

Results

Corpus Analysis

The step of generating representative keywords for each subcorpus (refer to the *Corpus analysis* subsection) resulted in a list of 3382 keywords. Sorted by decreasing the frequency difference, the top words (with *FreqDiff*>0) are representative of the suicidal case subcorpus (2360 keywords). Similarly, sorting in ascending order, top words (with *FreqDiff*<0) are representative of the nonsuicidal control subcorpus (1022 keywords). Table 4 shows the 10 top-ranking keywords for each subcorpus. In this table, we show each word's rank as well as its frequency in the whole corpus, the frequency difference between case and control subcorpora, and the frequency ratio for the word across the subcorpora. We provide a similar list of the top 100 keywords in Multimedia Appendix 2.

Table 4. Ranked keyword list for suicidal case and nonsuicidal control subcorpora.

Suicidal case keywords					Nonsuicidal control keywords				
Rank	Word	Freq[a] (words per million)	Freq diff[b]	Freq ratio[c]	Rank	Word	Freq (words per million)	Freq diff	Freq ratio
1	QQQQQ[d]	9779.1	3545.7	1.6	1	ZZZZZ[d]	35657.1	−3801.4	1.1
2	self	4278.5	2060.9	1.9	2	mental	3092.5	−1242.5	1.4
3	harm	2916.2	1673.4	2.4	3	mr	1197.9	−1138.1	2.0
4	ward	5554.7	1597.1	1.4	4	appointment	1583.5	−1124.5	1.7
5	overdose	1717.0	1392.8	5.3	5	medication	3756.5	−1017.4	1.3
6	staff	5670.0	1389.4	1.3	6	health	2282.2	−771.1	1.3
7	suicidal	2072.5	1256.2	2.5	7	please	1305.9	−703.6	1.5
8	said	5725.4	1137.7	1.3	8	state	1640.3	−694.4	1.4
9	alcohol	2276.2	1102.4	1.9	9	service	1190.6	−678.1	1.6
10	a&e	1534.1	1089.5	3.5	10	road	729.3	−596.2	1.8

[a]Freq: word frequency.

[b]Freq diff: frequency difference.

[c]Freq ratio: frequency ratio between subcorpora.

[d]Masking strings created by the electronic health record deidentification process: QQQQQ for relative or close contact identifiers and ZZZZZ for patient identifiers.

For the suicidal case subcorpus, the top keyword "QQQQQ" is a placeholder for anonymized names of relatives or close contacts of the patient created by a bespoke deidentification algorithm used in CRIS [45]. This could indicate concerns of relatives or carers being reported to staff over the patient's status. Other top keywords directly relate to the theme of suicide attempts (*overdose*, *suicidal*, and *a&e* [accident and emergency]). The frequency ratio indicates that *overdose* is over 5 times and *a&e* is over 3.5 times more frequent in the case subcorpus than in the control subcorpus. Other words relate to hospitalization (*ward* and *staff*) and self-harm (*self* and *harm*).

Visual inspection shows that *self* and *harm* frequently co-occur in noun phrases such as *harm to self* and *self-harm* (which was incorrectly segmented into 2 tokens by the tokenizer). Furthermore, *harm* also occurs with reflexive pronouns, for example, *harm himself/herself*, also referencing self-harm events. *Alcohol* is also clinically relevant because both chronic alcohol use disorders and acute use of alcohol confer risk for attempted suicide.

In contrast, for the control subcorpus, the top keyword "ZZZZZ" is a placeholder for anonymized patient identifiers. These top keywords are more generic terms that may be found in most types of clinical notes (eg, *mental*, *health*, and *state*) and some are likely to be derived from correspondence (eg, *mr*, *appointment*, and *please*). Although the top control keywords are significantly more frequent than those in the case subcorpus, the frequency difference and ratio are globally less marked than for case keywords. The median absolute frequency difference (*FreqDiff*) for the top 10 control keywords is 894.2, compared with 1391.1 for cases. The corresponding median frequency ratios (*FreqRatio*) are 1.90 for cases and 1.45 for controls. This indicates that keywords for suicide-related texts are more strongly representative of the case subcorpus than the keywords for the control subcorpus. This may reflect the fact that cases have a distinct unifying feature of being included for their hospitalized suicide attempt, whereas control admissions were from any period as long as they did not precede a suicide attempt. It should be noted that no suppositions about the sentiment associated with these keywords were made.

Sentiment Lexicon Analysis

We first assessed the global coverage of sentiment lexicons (refer to Multimedia Appendix 1 (E) for details). The figures for global coverage are summarized in Table 5.

Table 5. Term type and token counts for each lexicon in case and control subcorpora and whole corpus. Percentages for control words are shown as (raw/adjusted). Figures are in descending order of lexicon (filtered) size.

Lexicon	Filtered size	Word types			Word tokens		
		Case, n (%)	Control, n (%)	Whole corpus, n (%)	Case, n (%)	Control, n (%)	Whole corpus, n (%)
SentiWord-Net	39,746	9843 (9.02)	12,429 (7.64/5.12)	13,373 (6.46)	4,234,058 (50.49)	8,603,932 (65.19/41.42)	12,837,990 (59.48)
Opinion	6789	3111 (2.85)	3662 (2.25/1.51)	3821 (1.85)	979,804 (11.68)	1,959,007 (14.84/9.43)	2,938,811 (13.62)
EmoLex	5555	3733 (3.42)	4260 (2.62/1.75)	4426 (2.14)	1,456,097 (17.36)	2,869,472 (21.74/13.81)	4,325,569 (20.04)
AFINN	3478	2529 (2.32)	2781 (1.71/1.15)	2845 (1.37)	1,274,283 (15.20)	2,532,261 (19.19/12.19)	3,806,544 (17.64)
Pattern	2293	1101 (1.01)	1243 (0.76/0.51)	1296 (0.63)	910,369 (10.86)	1,957,386 (14.83/9.42)	2,867,755 (13.29)
LIWC[a]	1371	3708 (3.40)	5824 (3.58/2.40)	6269 (3.03)	620,546 (7.40)	1,830,216 (13.87/8.81)	2,450,762 (11.35)

[a]LIWC: Linguistic Inquiry and Word Count.

SentiWordNet, by far the largest lexicon, has the widest coverage of approximately 60% of all tokens (6.46% types) in the entire corpus. The pattern has the lowest word-type coverage for both subcorpora and the whole corpus (0.63%). Although LIWC has the fewest lexical entries (1371), its use of regular expressions that capture multiple word forms means it maps more individual word types (but has the lowest coverage of tokens, 11.35% on the whole corpus). Despite having approximately 1200 and 3300 fewer entries than Opinion, respectively, EmoLex and AFINN both have a substantially higher coverage of word tokens over the larger lexicon. EmoLex also has a slightly higher coverage of token types. This may be a consequence of the manner in which these lexicons were constructed and the sources from which they were derived. We review this issue in the *Discussion* section.

With the exception of LIWC, all lexicons show higher coverage of word types in the case subcorpus than in the control subcorpus. The same trend was observed when considering the adjusted percentages for word tokens. This suggests that there is generally more *sentiment* (as defined in these lexicons) expressed in the case subcorpus than in the control subcorpus, assuming an artificial scenario in which there are an equal number of words of each. However, if no adjustment for word frequency disparities across subcorpora is made, the opposite tendency is observed for all lexicons.

This notion of coverage does not take into account the representativeness of the words in question. To capture this crucial characteristic, we examined the proportion of keywords (word types) from each subcorpus containing each lexicon (keyword coverage; refer to the *Corpus Analysis* subsection and Multimedia Appendix 1 [D]). The overall proportional coverage of keywords is shown in Table 6.

Table 6. Case and control keywords that appear in each sentiment lexicon, in descending order of lexicon (filtered) size. The total number of keywords for the case subcorpus is 2360 and for the control subcorpus is 1022.

Lexicon	Filtered size	Case, n (%)	Control, n (%)
SentiWordNet	39,746	604 (25.6)	231 (22.6)
Opinion	6789	192 (8.1)	60 (5)
EmoLex	5555	277 (11.7)	117 (11.4)
AFINN	3478	238 (10.1)	74 (7)
Pattern	2293	115 (4.9)	39 (3)
LIWC[a]	1371	181 (7.7)	48 (4)

[a]LIWC: Linguistic Inquiry and Word Count.

As with global coverage, keyword coverage is correlated with lexicon size, with LIWC being the exception. Again, when examining only the most representative words for each subcorpus, Opinion, the second largest resource, has substantially lower coverage than both EmoLex and AFINN, which are smaller in size, the latter resource numbering only half as many keywords among its entries.

Evaluating the lexicons from an information retrieval perspective revealed the extent to which each lexicon strikes a balance between the inclusion of case keywords and the exclusion of

control keywords, accounting for the representativeness of the words identified. As shown in Table 7, all lexicons provided reasonable weighted precision (0.53-0.72). However, weighted recall and weighted F-score, which varied substantially across lexicons, were very low (0.04-0.36).

Table 7. Weighted metrics for each lexicon in descending order of weighted F score.

Lexicon	Weighted precision	Weighted recall	Weighted *F* score
SentiWordNet	0.68	0.36	0.47
EmoLex	0.68	0.18	0.29
AFINN	0.72	0.15	0.25
Opinion	0.68	0.11	0.18
LIWC[a]	0.69	0.10	0.17
Pattern	0.53	0.04	0.07

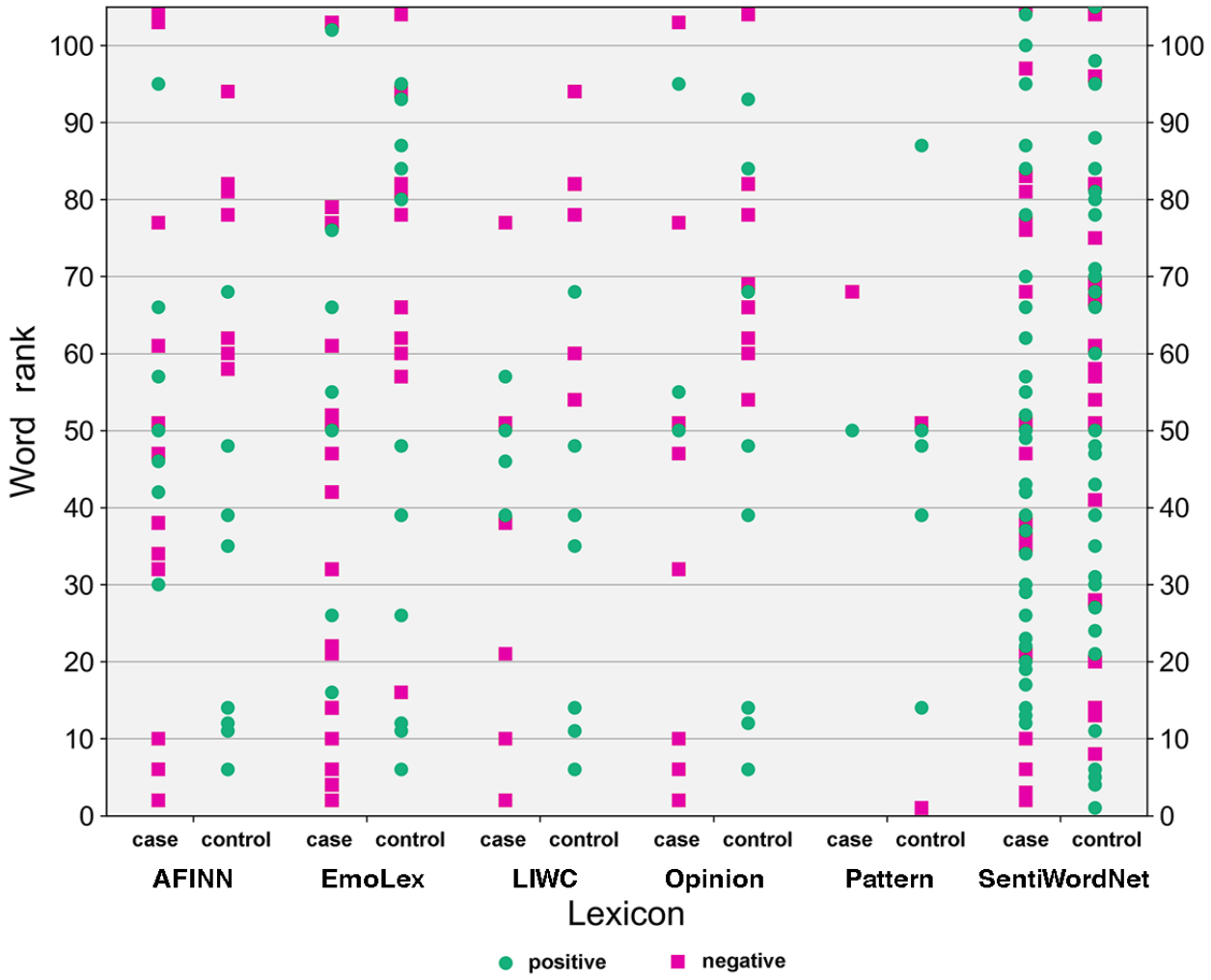
[a]LIWC: Linguistic Inquiry and Word Count.

These results show that, of all the lexicons we tested, SentiWordNet provides the best balance between precision and recall over keywords from the 2 subcorpora. Owing to its size, it obtained the highest recall. This indicates that it contains more of the most highly ranked case keywords than the other lexical resources. It also achieved precision on par with the other lexicons, indicating that the words it identifies are often high-ranking keywords from the suicide-related case subcorpus. The pattern lexicon achieved significantly lower results in terms of weighted precision and recall than all other lexicons, despite being larger than some of these. This suggests that its included sentiment terms are of a somewhat different nature and do not contribute a clear signal for distinguishing representative case keywords from control keywords.

Overall, as tools for distinguishing suicide-related from nonsuicide-related clinical notes, this evaluation, in particular the recall figures, shows that the most representative keywords in both subcorpora are not sentiment bearing, as defined in all these lexicons, thus indicating that there is a need for further analysis of the representative subcorpus keywords to better understand their characteristics.

Finally, we examined the distribution of sentiment among the top-ranking representative keywords for each subcorpus (sentiment coverage). Figure 1 shows the ranks of the top 100 keywords each lexicon contains for the case and control subcorpora. In addition to plotting the ranks of words featured in each lexicon, we also indicate, through color and shape coding, the polarity associated with each term.

Figure 1. Comparative sentiment lexicon coverage of top 100 ranked words for the suicidal case and nonsuicidal control subcorpora.



In terms of sentiment coverage, AFINN, EmoLex, LIWC, and Opinion mark a clear distinction between the top case and control keywords. These lexicons assign negative sentiment to high-ranking case keywords (eg, *harm* [ranked third], *risk* [11th], *kill* [52nd], and *pain* [78th]) and positive sentiment to top control keywords (eg, *please* [seventh], *calm* [40th], and *pleasant* [49th]), and negative also to certain high-ranking control keywords (eg, *aggressive* [61st], *illness* [63rd], and *anxiety* [83rd]).

Only 2 high-ranking keywords for cases appeared in the Pattern lexicon: these were *safe* [51st], which was the only one of the top 100 ranked words consistently found for cases across all 7 lexicons, and *past* [68th], which only appeared in Pattern and was ascribed a negative polarity (further discussed in the *Discussion* section). *Calm* [40th] and *pleasant* [49th] were the only top 100 keywords found consistently for controls across all 6 lexicons, and these were ascribed a positive polarity by all except SentiWordNet. This unexpected assignment of sentiment (the adjective *calm* is given a heavily negative score in SentiWordNet, whereas *anxious*, *borderline*, *cutting*, and *concern* are positive) highlights the importance of studying the underlying assumptions in off-the-shelf tools and their potential implications when applying them for a new use case.

For SentiWordNet, sentiment of top keywords is mixed, with a higher proportion of positive sentiment keywords in both subcorpora, although it assigned more negative sentiment for controls and for a greater proportion of the high-ranked keywords. This shows that despite having a larger lexical coverage, the sentiment coverage of this lexicon may not be sufficiently consistent to reliably distinguish the 2 populations.

It is important to note that 51 of the top 100 keywords for the case subcorpus were not identified by any of the lexicons. These included *self*, *staff*, *said*, *alcohol*, and *a&e*, all in the top 10 (Table 4), as well as further highly clinically relevant (although not necessarily sentiment bearing) words such as *paracetamol* (ranked 25th, FreqDiff=524.6, FreqRatio=4.5), the abbreviation *od* (used variably in psychiatry to mean either *overdose* or *omne in die* [once a day] with respect to medication; ranked 29th, FreqDiff=498.2, FreqRatio=2.2), *ambulance* (ranked 57th, FreqDiff=340.9, FreqRatio=3.3), the plural form *overdoses* (ranked 68th, FreqDiff=314.0, FreqRatio=7.6), and the acronym *dsh* (deliberate self-harm; ranked 83rd, FreqDiff=275.1, FreqRatio=3.4). The frequency ratio of these words shows that they were many times more frequent in suicide-related case notes than in the control corpus. Over the entire list of case keywords, only 33.35% (787/2360) were assigned a sentiment value by at least one of the lexicons. Furthermore, 51 of the top

100 control keywords were also absent from all lexicons, many of which pertain to correspondence (eg, *mr*, *appointment*, and *fax*). We refer the reader to Multimedia Appendix 2 for further details.

Discussion

Implications for Suicide Risk Assessment Lexicon Development

The list of representative keywords extracted from our corpus shows that the notion of sentiment generally adopted in the field of NLP is not the most appropriate semantic category for identifying terms that typify case notes of suicidal patients. Many of these terms do not carry an obvious negative or positive polarity, as defined in the tested sentiment lexicons.

Our analysis also showed that there is a need for further analysis of the assignment of sentiment polarity by these tools when applied on new use cases.

Furthermore, many of the keywords we identified as representative of suicide-related case notes were *neutral* with respect to sentiment, which is expected, and representative case keywords extracted in our study indicate that they are distinct from control keywords, but not all such terms would necessarily be sentiment bearing.

Our results show that these sentiment lexicons built using validated lexical resources, such as dictionaries or thesauri (eg, EmoLex), had higher combined precision and recall results than those derived from semiautomatic processes over large open-domain text corpora (eg, Opinion, built by web crawling).

Guidelines for Building Sentiment Lexicons for Suicide Risk Assessment

Following the work of Deng et al [24], one solution to the unsuitability of general-domain lexical resources for the clinical domain consists of defining the notion of sentiment for the analysis of clinical texts, and in the present case, of mental health (Guideline 1). This could allow the assignment of polarity to terms that do not feature in general-purpose lexical resources. In the case of suicide risk assessment, this might include the assignment of negative polarity to terms such as *a&e*, *overdose*, *alcohol*, *dsh*, and *plan*, which were not assigned a polarity value by the lexicons we tested.

In light of our results, a suggested strategy for building a suicide risk assessment lexicon may be to use corpus word frequencies as a guide to inclusion of words in a lexical resource that would remain agnostic with respect to sentiment (Guideline 2) and instead labeling terms as *trigger* or *risk factor* words (Guideline 3). Such a strategy would avoid the problem of assigning sentiment to words which, although highly representative of suicide-related texts, do not have an obvious *sentiment* value. This would also obviate the need to assign a polarity to terms that may be ambiguous in the sentiment they express, being either positive or negative depending on context (eg, *low* [emotion] vs *low* [risk]), although the more general problem of polysemy remains.

For clinically relevant terms, specialized psychiatric dictionaries or health care terminologies could be beneficial in creating a targeted lexical resource for suicide risk assessment (Guideline 4). For example, certain risk factors for suicide (eg, previous suicide attempts, depression, and substance misuse) and protective factors (eg, effective clinical care, family, and community support) are already well-known clinical features. Therefore, these concepts and associated terms should be reflected in any lexicon aiming to identify periods of increased suicide risk in clinical notes. One caveat that must be kept in mind is that many terms contained in specialized clinical terminologies are not written in EHRs by clinicians [46], meaning that term selection should be carried out by domain experts with a general awareness of typical target corpora.

Automated approaches to extracting terms from large corpora have become common in the field of NLP, including the creation of sentiment lexicons [47-49]. These techniques provide a means to increase the coverage of relevant terms, although it is preferable to implement some mechanism to ensure that the criterion of relevance is respected. Incorporating a domain-specific corpus-based notion of term *representativeness* into automatic lexicon induction procedures [50] is one way of refining term selection, filtering out terms that are deemed to be nonrepresentative (Guideline 5). Furthermore, a manual validation by domain experts (Guideline 6), where feasible, would further serve to ensure the precision of the extracted terms and could also be used to assign additional semantic categories such as sentiment.

Summary of guidelines is as follows:

1. Define the notion of sentiment for the clinical domain
2. Use corpus word frequencies as a guide to inclusion of words in a lexicon
3. Label terms as *risk factor* or *trigger* rather than sentiment-bearing
4. Use specialized dictionaries and/or health care terminologies as a source
5. Incorporate domain-specific corpus-based notion of representativeness into automatic lexicon induction techniques
6. Manual validation by domain experts

Summary and Limitations

Examining our data using the methods of corpus linguistics revealed statistically significant differences between the keywords used in EHR notes preceding an admission for attempted suicide and those from control periods not associated with such an attempt. Themes included hospitalized suicide attempts, self-harm, and alcohol. Coverage of these keywords by the general-purpose sentiment lexicons we reviewed was varied. Although lexicon size was a determining factor in overall coverage, the largest resource, SentiWordNet, did not distinguish the 2 subcorpora as well as some of the smaller resources, namely, AFINN, EmoLex, and Opinion, once both keyword rankings and sentiment were taken into consideration. Similarly, EmoLex and AFINN had wider coverage of relevant keywords than Opinion, which is the largest of the 3 resources. This may be partly a consequence of the original sampling strategy used to select words to construct sentiment lexicons. Both EmoLex

and AFINN were built on top of existing general-purpose dictionaries, whereas Opinion was created semiautomatically by crawling product reviews on the internet. As a result, the vocabulary of the latter may be more specific to that domain, whereas the 2 former lexicons are likely to be more generic in their terminology, meaning they may adapt slightly better to different domains. The same 3 lexicons also showed the most discriminating assignment of sentiment polarity between the case and control keywords. Although many of the terms contained in these resources can be said to convey appropriate sentiment values (eg, *anxiety* is negative and *pleasant* is positive), there are also certain terms for which this is less obvious, at least in the context of EHR text related to suicide risk. For example, *ward* is assigned negative sentiment by SentiWordNet, whereas *thoughts* are assigned positive sentiment. The word *plan* is assigned positive sentiment by EmoLex, whereas *call* is negative. Annotating word polarity in a noncontextual manner, especially without appropriate part-of-speech disambiguation (only 2 of the resources we tested contained entries with part-of-speech information), could lead to biased analyses in downstream modeling of new use cases. Clinical texts are intended to be written in an objective style, rather lacking what one might generally term *sentiment*, although in reality this may not always be the case. Many of the most highly relevant terms identified by our approach (eg, *a&e*, *overdoses*, and *alcohol*) do not fall into what might typically be termed a sentiment category but rather belong to categories of risk factors, whereas other identified terms are more sentiment bearing.

These observations lead us to concur with the conclusions of previous research [21-24] that domain-specific resources need to be developed for the analysis of clinical texts. We have attempted to provide insight into why this might be and what information such resources might need to include to address the task of suicide risk assessment through the analysis of clinical notes.

Our study has some limitations. First, the corpus was not constructed according to a deliberate sampling strategy but is the result of a 1:4 case-control selection ratio, which is typical in epidemiology. Completed and attempted suicide is much rarer than our sample suggests. Furthermore, the documents were not sampled according to type. This may have led to a preponderance of letters in the control corpus, as suggested by the most frequent keywords. The distribution of documents between patients also differs between the case and control subcorpora. Cases have, on average, almost 3 times the number of documents as controls, which is reflective of more frequent contact with mental health services. Consequently, the resulting corpus does not necessarily fulfill the criteria of representativeness and balance generally recommended in corpus linguistics.

We also acknowledge that our normalization of sentiment values for the sake of comparison does not necessarily reflect the actual quantity of sentiment assigned by all lexicons and invite the reader to refer to previous studies where *raw* sentiment scores are compared [20-22]. It is also worth noting that previous studies have shown that emotions, such as happiness expressed in social media posts, may vary with population demographics, geographical location [51,52], movement, and residency status in an area [53]. Although our work has focused on clinical texts instead of social media, such factors may have influenced our results; however, we have not controlled for this. This represents a caveat concerning the generalizability of our results to clinical populations in other geographical areas with potentially different sociodemographic configurations.

Finally, we only examined keywords that were common to both subcorpora. As a consequence, certain keywords typical of suicidal case notes only appearing in the case subcorpus may have been missed out, although we did find keywords appearing in only 1 subcorpus to be relatively infrequent compared with those we did examine.

Conclusions

This work makes several contributions to the study of sentiment in suicide risk assessment.

First, our corpus of clinical notes drawn from a case-control study of suicidal and nonsuicidal hospital admissions is, to our knowledge, a novel use of EHRs in this area.

Second, by applying methods of corpus linguistics, we identified 2 lists of keywords: the first representative of the clinical notes of patients leading up to a hospitalized suicide attempt and a second for those who made no such attempt. We used these lists of keywords to gauge the coverage of 6 sentiment lexicons over our corpus, using a number of measures, including information retrieval metrics, which we adapted for the purposes of our evaluation. Our study provided a novel examination of the content of these lexicons and their implications in relation to sentiment analysis as well as deeper insights into the characteristics of terms that distinguish suicide risk cases from controls in EHR text. Furthermore, we found that these general-domain resources assign polarity values that are sometimes not clinically meaningful or consistent with clinical judgments.

Finally, based on the outcomes of our study, we have suggested a set of simple and clear guidelines to facilitate the creation of more useful lexical resources for those seeking to assess risk of suicide through the analysis of clinical notes. Such targeted lexicons have the potential to advance research into the use of EHRs for the study of suicide risk in clinical populations by providing discriminative features for use in both rule-based and machine learning classification systems.

Acknowledgments

The authors wish to thank Jefrey Lijffijt and Paul Rayson for their advice on corpus linguistics and James Pennebaker for permission to use the LIWC lexicon. Any errors are the authors' own. RD is funded by a Clinician Scientist Fellowship (project e-HOST-IT) from the Health Foundation in partnership with the Academy of Medical Sciences, which also funds AB. This work was also

partly supported by Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England), and the devolved administrations, leading medical research charities, and the Maudsley Charity. This paper represents independent research partly funded (AR, RD, SV, and AB) by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

AB conceptualized the data curation, formal analysis, investigation, methodology, software, and writing–original draft preparation. SV conceptualized the methodology, and writing–review and editing. AR conceptualized and contributed in writing review and editing. RD conceptualized formal analysis, funding acquisition, supervision, writing–original draft preparation, and writing–review and editing.

Conflicts of Interest

RD and SV declare previous research funding received from Janssen.

Multimedia Appendix 1
Technical details of data and analyses.
[DOCX File , 102 KB - medinform_v9i4e22397_app1.docx]

Multimedia Appendix 2
Top 100 case and control keywords and associated polarities per lexicon.
[XLSX File (Microsoft Excel File), 31 KB - medinform_v9i4e22397_app2.xlsx]

References

1. Suicide data. World Health Organization. 2016. URL: http://www.who.int/mental_health/prevention/suicide/suicideprevent/en/ [accessed 2019-04-16]
2. Bostwick JM, Pabbati C, Geske JR, McKean AJ. Suicide Attempt as a Risk Factor for Completed Suicide: Even More Lethal Than We Knew. Am J Psychiatry 2016 Nov 01;173(11):1094-1100 [FREE Full text] [doi: 10.1176/appi.ajp.2016.15070854] [Medline: 27523496]
3. Franklin JC, Ribeiro JD, Fox KR, Bentley KH, Kleiman EM, Huang X, et al. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychol Bull 2017 Feb;143(2):187-232. [doi: 10.1037/bul0000084] [Medline: 27841450]
4. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front Psychiatry 2019 Feb 13;10:36 [FREE Full text] [doi: 10.3389/fpsyt.2019.00036] [Medline: 30814958]
5. Strauss J. Subjectivity and severe psychiatric disorders. Schizophr Bull 2011 Jan 20;37(1):8-13 [FREE Full text] [doi: 10.1093/schbul/sbq116] [Medline: 20961994]
6. Pang B, Lee L. Opinion mining and sentiment analysis. FNT in Information Retrieval 2008;2(1–2):1-135. [doi: 10.1561/1500000011]
7. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge, UK: Cambridge University Press; 2015.
8. Turney PD. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL'02.: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ?02. Association for Computational Linguistics; 2001 Presented at: The 40th Annual Meeting on Association for Computational Linguistics - ACL'02; July 2002; Philadelphia, PA, USA p. 417-424. [doi: 10.3115/1073083.1073153]
9. Hu M, Liu B. Mining and Summarizing Customer Reviews. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, United States: Association for Computing Machinery; 2004 Presented at: The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 2004; Seattle, WA, USA p. 168-177 URL: https://doi.org/10.1145/1014052.1014073 [doi: 10.1145/1014052.1014073]
10. Blair-Goldensohn S, Hannan K, McDonald R. Building a Sentiment Summarizer for Local Service Reviews. In: Proceedings of NLP in the Information Explosion Era (NLPIX 2008). 2008 Presented at: NLP in the Information Explosion Era (NLPIX 2008); April 2008; Beijing, China URL: https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/34368.pdf

11. Socher R, Perelygin A, Wu J. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. : Association for Computational Linguistics; 2013 Presented at: Conference on Empirical Methods in Natural Language Processing. Published online ?1642; 2013; Seattle, WA, USA p. 1631-1642.
12. Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: Proceedings of LREC 2010. Luxembourg: European Language Resources Association (ELRA); 2010 Presented at: Language Resources and Evaluation Conference (LREC 2010); May 2010; Valleta, Malta p. 1320-1326.
13. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R. Sentiment Analysis of Twitter Data. In: Proceedings of the Workshop on Languages in Social Media. Stroudsburg, PA, USA: Association for Computational Linguistics; 2011 Presented at: Workshop on Languages in Social Media (LSM'11); June 2011; Portland, OR, USA p. 30-38.
14. Dini L, Bittar A. Emotion Analysis on Twitter: the Hidden Challenge. In: Proceedings of LREC 2016. Luxembourg: European Language Resources Association (ELRA); 2016 Presented at: Language Resources and Evaluation Conference (LREC 2016); May 2016; Reykjavik, Iceland p. 3953-3958.
15. Smith P, Lee M. Cross-discourse Development of Supervised Sentiment Analysis in the Clinical Domain. Stroudsburg, PA, USA: Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics; 2012 Presented at: Third Workshop in Computational Approaches to Subjectivity and Sentiment Analysis; July 2012; Jeju, Korea p. 79-83.
16. Xia C, Zhao D, Wang J, Liu J, Ma J. ICSH 2018: LSTM based Sentiment Analysis for Patient Experience Narratives in E-survey Tools. In: Chen H, Fang Q, Zeng D, Wu J, editors. Proceedings of ICSH 2018. Berlin/Heidelberg, Germany: Springer International Publishing; 2018:231-239.
17. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. Journal of Medical Internet Research 2013;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
18. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In: Li J, Cao L, Wang C, Tan KC, Liu B, Pei J, et al, editors. Trends and Applications in Knowledge Discovery and Data Mining. Berlin/Heidelberg, Germany: Springer; 2013:201-213.
19. Tao X, Zhou X, Zhang J, Yong J. Sentiment Analysis for Depression Detection on Social Networks. In: Li J, Li X, Wang S, Li J, Sheng QZ, editors. Advanced Data Mining and Applications. Berlin/Heidelberg, Germany: Springer International Publishing; 2016:807-810.
20. McCoy TH, Castro VM, Cagan A, Roberson AM, Kohane IS, Perlis RH. Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study. PLoS ONE 2015 Aug 24;10(8):e0136341. [doi: 10.1371/journal.pone.0136341]
21. McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. JAMA Psychiatry 2016 Oct 1;73(10):1064-1071. [doi: 10.1001/jamapsychiatry.2016.2172] [Medline: 27626235]
22. Weissman GE, Ungar LH, Harhay MO, Courtright KR, Halpern SD. Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness. J Biomed Inform 2019 Jan;89:114-121 [FREE Full text] [doi: 10.1016/j.jbi.2018.12.001] [Medline: 30557683]
23. Deng Y, Stoehr M, Denecke K. Retrieving Attitudes: Sentiment Analysis from Clinical Narratives. In: Proceedings of the Medical Information Retrieval (MedIR) Workshop. New York, NY, USA: Association for Computing Machinery; 2014 Jul 11 Presented at: 37th international ACM SIGIR conference on Research & development in information retrieval; July 2014; Gold Coast, Australia p. 12-15.
24. Deng Y, Declerck T, Lendvai P, Denecke K. The Generation of a Corpus for Clinical Sentiment Analysis. In: Sack H, Rizzo G, Steinmetz N, Mladenić D, Auer S, Lange C, editors. Lecture Notes in Computer Science. Cham, Switzerland: Springer International Publishing; 2016:311-324.
25. Holderness E, Cawkwell P, Bolton K, Pustejovsky J, Hall MH. Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Stroudsburg, PA, United States: Association for Computational Linguistics; 2019 Presented at: 2nd Clinical Natural Language Processing Workshop; June 2019; Minneapolis, MN, USA p. 117-123. [doi: 10.18653/v1/w19-1915]
26. Perera G, Broadbent M, Callard F, Chang C, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) Case Register: current status and recent enhancement of an Electronic Mental Health Record-derived data resource. BMJ Open 2016 Mar 1;6(3):e008721 [FREE Full text] [doi: 10.1136/bmjopen-2015-008721] [Medline: 26932138]
27. Grimes DA, Schulz KF. Compared to what? Finding controls for case-control studies. Lancet 2005 Apr;365(9468):1429-1433. [doi: 10.1016/s0140-6736(05)66379-9]
28. spaCy · Industrial-strength Natural Language Processing in Python. URL: https://spacy.io/ [accessed 2020-07-07]
29. Bird S, Klein E, Loper E, Baldridge J. Multidisciplinary instruction with the Natural Language Toolkit. In: Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL'08). Stroudsburg, PA, USA: Association for Computational Linguistics; 2008 Presented at: Third Workshop on Issues in Teaching Computational Linguistics (TeachCL'08); June 2008; Columbus, OH, USA p. 62-70. [doi: 10.3115/1627306.1627317]

30. KCL-Health-NLP/Suicide-Risk-Sentiment. URL: https://github.com/KCL-Health-NLP/suicide-risk-sentiment [accessed 2020-11-09]
31. Kilgarriff A. Comparing Corpora. IJCL 2001 Dec 17;6(1):97-133. [doi: 10.1075/ijcl.6.1.05kil]
32. Paquot M, Bestgen Y. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. In: Jucker AH, Schreier D, Hundt M, editors. Pragmatics and Discourse. Leiden, Netherlands: Brill; 2009.
33. Lijffijt J, Nevalainen T, Säily T, Papapetrou P, Puolamäki K, Mannila H. Significance testing of word frequencies in corpora. Digital Scholarship Humanities 2014 Dec 8;31(2):374-397. [doi: 10.1093/llc/fqu064]
34. Gries ST. Dispersions and adjusted frequencies in corpora. IJCL 2008;13(4):403-437. [doi: 10.1075/ijcl.13.4.02gri]
35. Nielsen F. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs.: arXiv; 2011. URL: http://arxiv.org/abs/1103.2903 [accessed 2019-02-25]
36. Mohammad S, Turney P. Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence 2013;29(3):436-465. [doi: 10.1111/j.1467-8640.2012.00460.x]
37. Tausczik YR, Pennebaker JW. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology 2009 Dec 08;29(1):24-54. [doi: 10.1177/0261927X09351676]
38. De Smedt T, Daelemans W. Pattern for Python. J Mach Learn Res 2012;13:2063-2067.
39. Baccianella S, Esuli A, Sebastiani F. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: Proceedings of the 7th edition of the Language Resources and Evaluation Conference (LREC 2010). 2010 Presented at: Language Resources and Evaluation Conference (LREC 2010); 2010; Valleta, Malta p. A.
40. Miller GA. WordNet: a lexical database for English. Commun ACM 2019;38(11):39-41. [doi: 10.1145/219717.219748]
41. Bernard J. The Macquarie Thesaurus. In: Macquarie Library. Sydney, Australia: Macquarie Library; 1986.
42. Strapparava C, Valitutti A. WordNet-Affect: an Affective Extension of WordNet. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004). 2004 Presented at: Language Resources and Evaluation Conference (LREC 2004); May 2004; Lisbon, POrtugal.
43. Stone P, Dunphy D, Smith M, Ogilvie D. The General Inquirer: A Computer Approach to Content Analysis. Cambridge, MA, USA: MIT Press; 1966.
44. Brants T, Franz A. Web 1T 5-Gram Version 1.: Linguistic Data Consortium; 2006. URL: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 [accessed 2019-04-20]
45. Fernandes AC, Cloete D, Broadbent MT, Hayes RD, Chang C, Jackson RG, et al. Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records. BMC Med Inform Decis Mak 2013 Jul 11;13:71 [FREE Full text] [doi: 10.1186/1472-6947-13-71] [Medline: 23842533]
46. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA. Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics 2010 Mar 31;1(1):5 [FREE Full text] [doi: 10.1186/2041-1480-1-5] [Medline: 20618981]
47. Tai YJ, Kao HY. Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation. In: Proceedings of International Conference on Information Integration and Web-Based Applications & Services (IIWAS'13). New York, NY, USA: Association for Computing Machinery; 2013 Presented at: International Conference on Information Integration and Web-Based Applications & Services (IIWAS'13); December 2013; Vienna, Austria p. 53-62. [doi: 10.1145/2539150.2539190]
48. Hamilton W, Clark K, Leskovec J, Jurafsky D. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora.: arXiv URL: http://arxiv.org/abs/1606.02820 [accessed 2019-04-19]
49. Yang X, Zhang Z, Zhang Z, Mo Y, Li L, Yu L, et al. Automatic construction and global optimization of a multisentiment lexicon. Comput Intell Neurosci 2016;2016:2093406-2093408 [FREE Full text] [doi: 10.1155/2016/2093406] [Medline: 28042290]
50. Pryzant R, Shen K, Jurafsky D, Wagner S. Deconfounded Lexicon Induction for Interpretable Social Science. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics; 2018 Presented at: Human Language Technologies; June 2018; New Orleans, LA, USA p. 1615-1625. [doi: 10.18653/v1/n18-1146]
51. Mitchell L, Frank MR, Harris KD, Dodds PS, Danforth CM. The geography of happiness: connecting twitter sentiment and expression, demographics, and objective characteristics of place. PLoS One 2013 May 29;8(5):e64417 [FREE Full text] [doi: 10.1371/journal.pone.0064417] [Medline: 23734200]
52. Gore RJ, Diallo S, Padilla J. You are what you tweet: connecting the geographic variation in America's obesity rate to twitter content. PLoS One 2015;10(9):e0133505 [FREE Full text] [doi: 10.1371/journal.pone.0133505] [Medline: 26332588]
53. Padilla JJ, Kavak H, Lynch CJ, Gore RJ, Diallo SY. Temporal and spatiotemporal investigation of tourist attraction visit sentiment on Twitter. PLoS One 2018 Jun 14;13(6):e0198857 [FREE Full text] [doi: 10.1371/journal.pone.0198857] [Medline: 29902270]

Abbreviations

CRIS: Clinical Record Interactive Search
EHR: electronic health record

Freq: frequency
FreqDiff: frequency difference
FreqRatio: frequency ratio
ICD: International Classification of Diseases
LIWC: Linguistic Inquiry and Word Count
NLP: natural language processing

Edited by C Lovis; submitted 10.07.20; peer-reviewed by S Mohammad, R Gore; comments to author 20.09.20; revised version received 26.11.20; accepted 05.12.20; published 13.04.21.

Please cite as:
Bittar A, Velupillai S, Roberts A, Dutta R
Using General-purpose Sentiment Lexicons for Suicide Risk Assessment in Electronic Health Records: Corpus-Based Analysis
JMIR Med Inform 2021;9(4):e22397
URL: https://medinform.jmir.org/2021/4/e22397
doi:10.2196/22397
PMID:33847595

Original Paper

Physician Stress During Electronic Health Record Inbox Work: In Situ Measurement With Wearable Sensors

Fatema Akbar[1], MSc; Gloria Mark[1], PhD; Stephanie Prausnitz[2], MS; E Margaret Warton[2], MPH; Jeffrey A East[3,4,5], MD, MPH; Mark F Moeller[3,6], MD; Mary E Reed[2], DrPH; Tracy A Lieu[2,3], MD, MPH

[1]Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA, United States

[2]Division of Research, Kaiser Permanente Northern California, Oakland, CA, United States

[3]The Permanente Medical Group, Oakland, CA, United States

[4]Department of Adult and Family Medicine, Kaiser Permanente, Richmond, CA, United States

[5]Department of Adult and Family Medicine, Kaiser Permanente, San Rafael, CA, United States

[6]Department of Adult and Family Medicine, Kaiser Permanente, Napa, CA, United States

Corresponding Author:
Tracy A Lieu, MD, MPH
Division of Research
Kaiser Permanente Northern California
2000 Broadway
Oakland, CA, 94611
United States
Phone: 1 510 891 3407
Email: tracy.lieu@kp.org

Abstract

Background: Increased work through electronic health record (EHR) messaging is frequently cited as a factor of physician burnout. However, studies to date have relied on anecdotal or self-reported measures, which limit the ability to match EHR use patterns with continuous stress patterns throughout the day.

Objective: The aim of this study is to collect EHR use and physiologic stress data through unobtrusive means that provide objective and continuous measures, cluster distinct patterns of EHR inbox work, identify physicians' daily physiologic stress patterns, and evaluate the association between EHR inbox work patterns and physician physiologic stress.

Methods: Physicians were recruited from 5 medical centers. Participants (N=47) were given wrist-worn devices (Garmin Vivosmart 3) with heart rate sensors to wear for 7 days. The devices measured physiological stress throughout the day based on heart rate variability (HRV). Perceived stress was also measured with self-reports through experience sampling and a one-time survey. From the EHR system logs, the time attributed to different activities was quantified. By using a clustering algorithm, distinct inbox work patterns were identified and their associated stress measures were compared. The effects of EHR use on physician stress were examined using a generalized linear mixed effects model.

Results: Physicians spent an average of 1.08 hours doing EHR inbox work out of an average total EHR time of 3.5 hours. Patient messages accounted for most of the inbox work time (mean 37%, SD 11%). A total of 3 patterns of inbox work emerged: inbox work mostly outside work hours, inbox work mostly during work hours, and inbox work extending after hours that were mostly contiguous to work hours. Across these 3 groups, physiologic stress patterns showed 3 periods in which stress increased: in the first hour of work, early in the afternoon, and in the evening. Physicians in group 1 had the longest average stress duration during work hours (80 out of 243 min of valid HRV data; P=.02), as measured by physiological sensors. Inbox work duration, the rate of EHR window switching (moving from one screen to another), the proportion of inbox work done outside of work hours, inbox work batching, and the day of the week were each independently associated with daily stress duration (marginal R^2=15%). Individual-level random effects were significant and explained most of the variation in stress (conditional R^2=98%).

Conclusions: This study is among the first to demonstrate associations between electronic inbox work and physiological stress. We identified 3 potentially modifiable factors associated with stress: EHR window switching, inbox work duration, and inbox work outside work hours. Organizations seeking to reduce physician stress may consider system-based changes to reduce EHR window switching or inbox work duration or the incorporation of inbox management time into work hours.

(JMIR Med Inform 2021;9(4):e24014) doi:10.2196/24014

KEYWORDS

electronic health records; stress; wearables; HRV; inbox; EHR alerts; after-hours work; electronic mail; physician well-being; Inbasket

Introduction

Background

Inbox management is an important component of electronic health record (EHR) work for physicians and a key potential stressor [1]. Through their EHR inbox, physicians receive messages from other physicians, staff, and patients. Studies of inbox management in other professions repeatedly report inbox management as a source of stress due to the time it takes to go through an ever-increasing volume of emails, the task demands associated with emails, and the interruptions they create [2-4]. Similarly, EHR inbox management has been identified as a possible contributor to physician stress and burnout [5,6]. To understand the relationship between EHR adoption and use and stress, it is critical to examine how physicians spend time on the EHR inbox.

Although several studies have addressed the stress or burden related to EHR use, there are two main limitations in previous work. First, scant research focusing on the inbox component of the EHR exists [1,5,7,8]. Second, previous studies relied on self-reported stress measured at a single time point (or a few time points) [5], which fails to capture the detailed and continual stress and EHR work patterns throughout the day and is prone to bias [9,10].

Our study investigates physicians' EHR inbox use patterns and associated stress, as measured unobtrusively and continuously by EHR system logs and wearable sensors. The objectives of this study are as follows:

1. Collect EHR use and stress data through unobtrusive means that provide objective and continuous measures.
2. Cluster and visualize distinct EHR inbox work patterns and identify their characteristics.
3. Identify physicians' daily stress patterns.
4. Evaluate the association between EHR inbox work characteristics and physician stress.

Previous Work on Physician Workload Related to the EHR and EHR Inbox

Studies have noted the burden of EHR digital work for physicians [11-13]. EHR-related factors that could lead to physician stress and burnout include the extra time needed, often beyond work hours, to complete EHR-related work [14-17], usability issues [18-20], risks associated with errors [21], and taking time out from face-to-face interactions with patients [22].

For EHR inbox management, a 2017 study [14] using EHR logs found that time spent in the inbox accounted for 24% of total EHR time, and of the time spent in the inbox, a larger proportion was spent after work hours compared with the time spent on other EHR activities. A study reported that 86% of surveyed physicians worked outside of work hours to respond to inbox messages [23], whereas another study reported that 37% of inbox work was done outside of work hours [24]. In addition to the time it takes within and outside of work hours, inbox-related burden has been attributed to the volume and source of EHR messages [5,7] and information overload from notifications (ie, asynchronous alerts) [25]. A 2012 study based on EHR logs [26] found that primary care physicians (PCPs) received a mean of 56.4 alerts per day and spent an estimated average of 49 minutes per day processing their alerts. A more recent study [1] found that PCPs received a mean of 77 (SD 38) inbox message notifications per day compared with the 30 notifications for specialists. Message quantity has been associated with increased attention switching and inbox work duration [27]. However, although these studies quantified EHR inbox–related factors and measured self-reported workload, well-being, or burnout at a single time point, they did not measure daily stress associated with EHR inbox use.

Unobtrusive Sensing of Stress

One of the main limitations of previous studies on EHR and stress is the reliance on self-reported measures of well-being and burnout collected at a single time point [7,18]. In addition to not directly measuring stress per se, self-report approaches have several limitations for stress monitoring in the workplace. When people subjectively report how they feel, their evaluation could be affected by memory bias and emotion recognition, regulation, and expression biases [9,10,28,29]. Administering surveys for self-reports can also be disruptive, as they require the full cognitive attention of the user and do not allow continuous or frequent measurement that could be correlated with inbox use.

Advances in wearable sensors and algorithms that filter and analyze their data enable objective, continuous unobtrusive sensing of physiological measures directly associated with stress, such as heart rate variability (HRV). HRV is the variation in time between one heartbeat and the next. When relaxing and recovering, HRV increases, and it decreases during stress [30-32]. Thus, measuring HRV throughout the day can provide an objective and continuous measure of stress and relaxation, which can be used to identify events associated with stress in more granularity than is possible with self-reports.

Compared with other physiological stress measures that can be obtained from wearable sensors in daily life, HRV is more reliable in real-world settings (outside the laboratory). For example, skin conductance (ie, electrodermal activity [EDA]) can be difficult to measure in dry, indoor air-conditioned settings as the electrodes rely on sweat to measure conductance. In addition, some people do not naturally produce adequate EDA signals [33]. HRV sensors in wrist-wearable devices are light based (photoplethysmography sensors) and are more commonly used in consumer-grade wearables.

HRV is affected by a number of factors other than stress, such as physical activity and overall health. Thus, HRV as a measure of stress is most reliable for healthy participants in sedentary settings. Previous studies used HRV from wearable devices as a measure of stress in office settings where participants were working on a computer [34-37], making this method applicable to computer-based work by physicians.

Methods

Study Setting

Data collection was conducted at one of the largest medical groups in the United States. The medical group has 9200 physicians and serves 4.4 million members in 21 hospital-based medical centers.

Since 2008, the participating medical group has been using a comprehensive EHR (Epic Systems) that integrates inpatient, emergency, and outpatient care, including primary care, specialty, laboratory, pharmacy, and imaging data. The EHR inbox, named the Inbasket, receives messages sent by patients via a portal website (also available through patient-facing mobile apps) and messages from other physicians, clinical staff, the pharmacy, laboratory, and other departments. Physicians can access the Inbasket on computers or mobile devices. Physicians are expected to respond to each patient message within 2 business days. Patients are encouraged to use the messaging functionality of the EHR to enhance access to their physicians and the care experience.

Typical work hours when clinical settings are open and patient appointments are booked are from 8:30 AM to 12:30 PM and 1:30 PM to 5:30 PM. Clinic time is dedicated to patient appointments, which are conducted in person in the clinic or via telephone or video telemedicine. Some physicians also do clinical work during weekends, with work hours that might differ from weekdays.

Recruitment and Protocol

Adult PCPs from 5 medical facilities within the medical group were recruited. Between 7 and 12 physicians were enrolled at each facility, with a total of 47 eligible physicians enrolled.

Physicians were eligible if they performed outpatient clinical work for at least 3.5 days a week. Physicians who were taking cardiac medications, had pacemakers or defibrillators, or had been diagnosed with cardiac arrhythmias were not eligible because of the interference of these factors with the HRV-based stress measure. Eligibility was confirmed via a recruitment email.

After obtaining written informed consent, the staff assigned a wearable device with heart rate sensors (Garmin Vivosmart 3) and configured the associated mobile apps (Garmin Connect and Tesserae Phone Agent [38]) on the physician's work-issued mobile phone. The apps streamed data from the wearable device via Bluetooth and uploaded the data to a server. The research team also installed an experience sampling app [39] on the physician's mobile phone to send short questions at specified times (see the *Experience Sampling* section). At enrollment, physicians completed a brief 5-question written survey about their EHR inbox management and stress.

Physicians were asked to wear the device and respond to the daily short survey prompts for 7 consecutive days and keep their phones and the wearable device charged. Physicians were free to keep their wearable devices after data collection. The study protocol was approved by the institutional review board of Kaiser Permanente Northern California.

Data

EHR System Logs

We used system access logs, which contained granular timestamped data on the Epic system EHR use. We created hourly time bins and variables from the log data to quantify how time was attributed to different activities and different types of inbox messages per hour. These variables, which were collected for every hour, included the number of minutes spent in the EHR, the number of minutes spent in the inbox, the number of minutes spent working on each inbox message type, the number of tasks performed, and the number of window switches (ie, clicking a new computer window).

We categorized the system-generated labels for message type description into high-level categories by analyzing the frequency of the labels along with input from our clinical collaborators who are familiar with the meanings and patterns of different types of messages. This approach resulted in 4 message types: (1) messages from patients; (2) results, such as laboratory test results; (3) requests, which ask the physician to perform an action such as approving a medication refill or signing clinical orders; and (4) informational and administrative messages. No message content or metadata (ie, sender, receiver, and message ID) were collected.

HRV-Based Measure of Stress

The device used to measure HRV (Garmin Vivosmart 3) was a wrist-worn device with an optical heart rate sensor. It produces a *stress score* based on HRV in still moments (ie, excluding times with physical activity that interfere with HRV readings) and accounts for the physiological norm of each user. The stress score ranges from 0 to 100 and is provided via the Garmin application programming interface as 3-minute averages of the real-time stress scores generated on the device. The stress analysis method used by the device has been empirically tested and validated [40]. Garmin heart rate sensors were also compared with other devices and were found to be among the most accurate devices [41-44].

In our analyses, the HRV-based stress measure was the duration (number of minutes) of medium and high stress (stress score of >50). We excluded low stress periods (scores from 25 to 50) because a certain amount of physiological stress indicates arousal which is expected (and needed) for performing daily tasks [45].

There were some gaps in the continuous HRV stress data (see the *Analysis* section). Missing HRV stress data could be attributed to loose fitting of the sensors on the wrist, removing the device for charging, or forgetting to wear the device or physical activity. We set a minimum of 20 minutes of HRV

data per hour for hourly stress measures and 2 hours of data for daily measures to be included in the analyses. We further report the number of valid minutes of data on which each reported stress measure is based.

Experience Sampling

During the data collection period, physicians received 3 short daily surveys via the experience sampling app. The survey consisted of 3 questions asking physicians to rate their stress in the last 5 minutes (from no stress to high stress), their arousal level (from low energy to high energy), and their mood (from unpleasant to pleasant). The experience sampling app triggered a phone notification asking physicians to take the survey 3 times a day: morning (between 9:30 AM and 10:30 AM), lunchtime (between 1 PM and 1:30 PM), and afternoon (between 3 PM and 4 PM). The survey expired 45 minutes after the notification if not opened.

Self-Reported Inbox Management Strategies and Related Stress

At enrollment, physicians were asked to complete a 5-question survey on their strategies for and feelings about Inbasket (their EHR inbox) management. Physicians were asked to indicate how distressful they found inbox management and whether they had responsibilities that restricted their ability to work before or after formal work hours.

Physician Characteristics

We also obtained physicians' age, sex, years of experience, and full-time equivalent (FTE) status, which is a measure of clinical workload where 40 hours per week of scheduled work is 1.0 FTE. According to internal analyses by the medical group, FTE is strongly correlated with the patient panel size for physicians.

Analysis

We used the Gaussian Mixture Models clustering algorithm [46] to find distinct patterns of inbox work. Features in the model included the distribution of inbox time in work hours and outside of work hours contiguous and noncontiguous to work hours. Multiple feature and cluster counts were tested, and the clustering that yielded more balanced clusters and had a reasonable silhouette score (a score that indicates how distinct or overlapping the clusters are) [47] was selected.

To capture whether physicians dedicated certain blocks of time for inbox work or consistently checked their inbox throughout the day, we defined days with inbox work batching as days where 70% or more of the total inbox work duration occurred in 3 separate blocks of time or less. With consistent inbox checking, a uniform distribution of inbox duration over the day would typically be observed, whereas batching would show 2-3 daily peaks of high inbox duration [35]. We compared this measure across clusters and used it as an independent variable in the mixed effects model along with the other EHR inbox use characteristics.

To compare clusters (ie, groups of different inbox work patterns), each comparison variable was tested for normality and homogeneity of variances before conducting an analysis of variance for normal distributions with equal variances or the Kruskal-Wallis test otherwise. For pairwise comparisons, a posthoc analysis was conducted using the Tukey honestly significant difference test for normally distributed variables and Dunn test for nonparametric posthoc comparisons. Categorical variables were tested using the Chi-square test.

To plot hourly stress patterns, we removed hours with less than 20 minutes of valid HRV data to avoid overestimating the stress duration as a ratio of the measurement period (the measurement period being valid HRV measurement duration). From a total of 4245 hours, this filter removed 1177 hours (27.73%) of the workdays' HRV data. For daily stress measures, workdays with less than 2 hours of valid HRV data were removed from the analysis, as well as workdays that are Saturdays or Sundays, and those with no inbox activity. This filter removed 21 days in total, keeping 178 workdays for the daily stress analyses (cluster comparison and a regression model).

We investigated the relationship between daily EHR inbox use and stress through a generalized mixed effects model with physicians as random effects. A Poisson distribution was used to represent stress minutes as events within the observation period (ie, valid HRV minutes as an offset in the model). The distribution of the dependent variable (ie, stress duration) was right skewed, as expected in a Poisson distribution. The independent variables were centered (ie, mean subtracted). The variance inflation factor was under 5 for all independent variables, indicating that multicollinearity was not a problem. Several models were compared, starting with a base model and incrementally adding variables, to ensure that the improvement in the model justified the added complexity of adding variables. The model with the lowest Akaike information criterion and highest marginal (fixed effects) R^2 is presented.

Results

Participants

The 47 physicians (32/47, 68% female) were aged an average of 43.83 years (SD 9.51; range 31-68), had an average of 15.17 (SD 9.93; range 4-42) years of experience in medicine, and had an average FTE of 81% (SD 14%). On average, physicians in the data set had 5.26 workdays (SD 0.94) and 2.74 nonworkdays (SD 0.94) over the 8 days of data collection (the day of enrollment plus 7 days in the study).

The HRV-based stress analyses included 42 physicians, because 5 physicians (1 male and 4 female) had technical issues, thereby causing loss of the wearable device data.

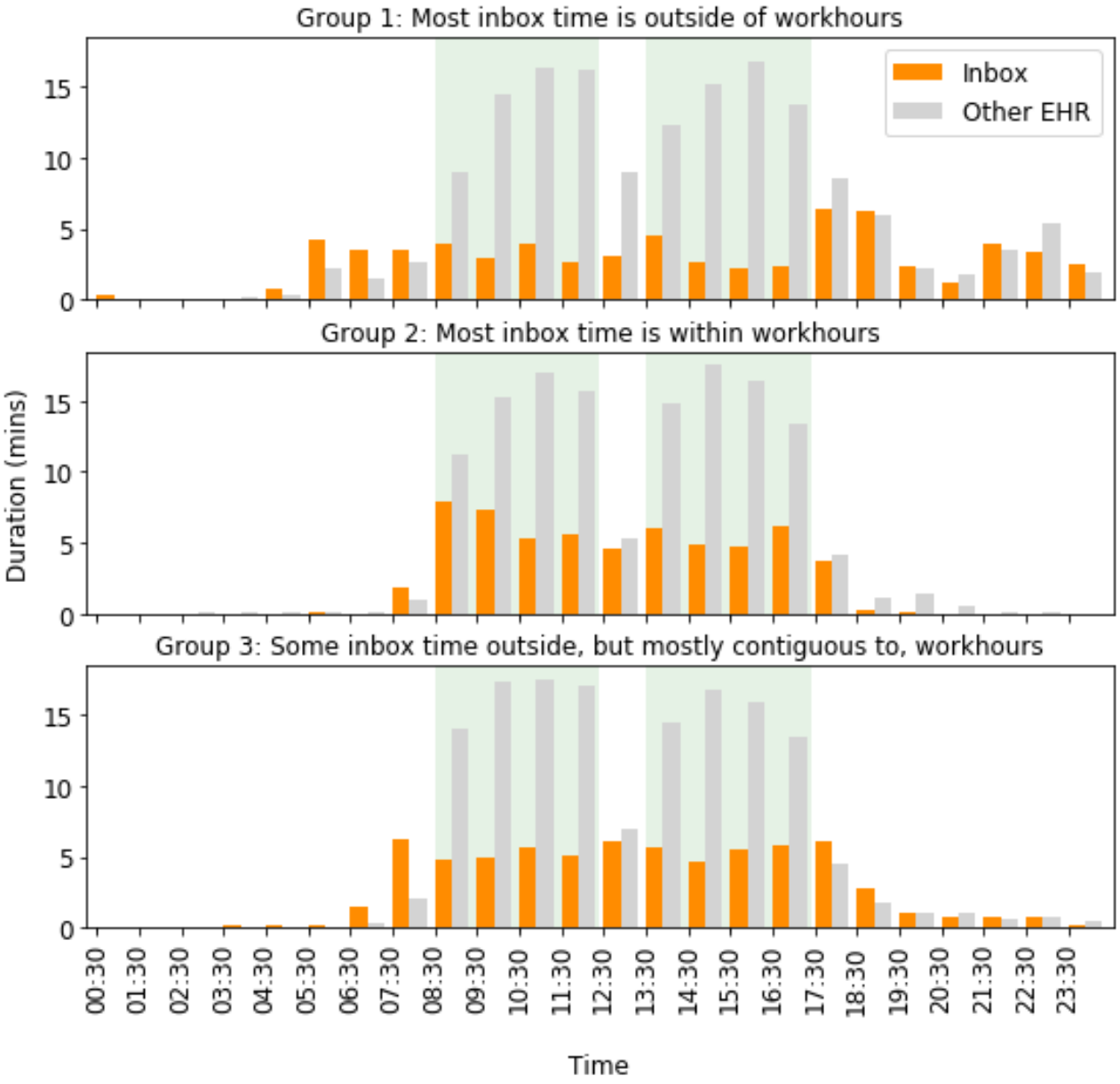
The inbox strategies and stress survey was completed by 44 physicians.

Three Distinct Patterns of EHR Inbox Work

On workdays, physicians spent an average of 3.5 hours (SD 0.69) in the EHR, of which 1.08 hours (SD 0.38) were spent doing inbox work. On nonworkdays, physicians spent an average of 23.88 minutes (SD 36.3) in the EHR, including an average of 13.78 minutes in inbox (SD 23.78). The majority of time in the inbox was spent on patient messages (mean 37%, SD 11%), followed by laboratory results (mean 31%, SD 8%), requests (mean 20%, SD 6%), and administrative messages (mean 13%, SD 5%).

Using the Gaussian Mixture Models clustering algorithm, we found 3 temporal patterns of work, with a silhouette score of 0.41, indicating moderate separation between these clusters (ie, distinct groupings). Figure 1 shows the average hourly time spent in the inbox and other EHR work (such as charting and order entry) for physicians in each cluster. Group 1 (n=10) represented physicians who spent time in the inbox outside work hours, in the evenings and early mornings; group 2 (n=17) represented physicians who worked mostly within work hours; and group 3 (n=20) represented physicians who spent some time on inbox work after hours that were mostly contiguous to work hours.

Figure 1. Temporal patterns of inbox and other EHR work. The green background indicates work hours. EHR: electronic health record.



Free-text responses from the survey on inbox management strategies supported these computationally generated inbox work patterns. Responses from physicians in group 1 indicated working beyond work hours, either by staying late in the office or taking work home. Some representative comments were as follows. A physician in group 1 reported, “I find when I sacrifice sleep to do more at home, I’m too tired during the day and I’m very inefficient at night,” indicating that they were working late at night. Physicians in group 2 indicated working mostly within work hours. For example, one physician in this group asserted, “I arrive around 8:30 and prefer to leave around 5:30.” Another stated: "I just like to work and finish work during my allotted work time. I do not like to work at other times or at home."

Physicians in group 3 also indicated not taking work home but at the cost of staying late in the office to clear their inbox. For example, a physician in group 3 said, “I generally try not to take work home [...] so often stay very late to clean out inbasket.”

Physician characteristics (age, sex, years of experience, and FTE) did not show statistically significant differences across the 3 work patterns. In terms of EHR use, total daily time spent on inbox work and other EHR work on workdays (24-hour period) did not differ across groups (P=.38 and P=.15,

respectively). However, as shown in Table 1, physicians in group 1 spent more time in the inbox after work hours compared with other groups, both in minutes and as a percentage of daily inbox time (P<.001). Posthoc comparisons showed that all the groups differed from each other. Group 1 also spent more time in the inbox work on nonworkdays (P=.03).

Table 1. Comparing inbox use characteristics across 3 work patterns.

Inbox use characteristics	Group 1, mean (SD)	Group 2, mean (SD)	Group 3, mean (SD)	*P* value
Clustering factors (percentage of all-day inbox duration)				
Work hours inbox duration	37 (12)	82 (8)	62 (9)	<.001
Outside and noncontiguous to work hours	42 (11)	1 (2)	12 (5)	<.001
Contiguous to work hours	21 (11)	17 (7)	26 (13)	.03
Duration of inbox work on workdays and nonworkdays (min)				
Work hours inbox duration	25.36 (13.03)	47.97 (13.35)	42.13 (16.56)	.002
Outside work hours inbox duration	41.37 (13.81)	10.91 (5.63)	26.97 (13.26)	<.001
Inbox duration on nonworkdays	32.74 (37.46)	11.13 (19.69)	6.54 (11.3)	.03
Message types (percentage of all inbox time)				
Patients	32 (10)	35 (10)	42 (10)	.02
Results	30 (9)	32 (11)	26 (10)	.10
Requests	24 (7)	20 (6)	21 (6)	.31
Admin	14 (5)	13 (4)	11 (4)	.14

Physicians in group 1 were more likely to batch their inbox work (ie, do most of their inbox work in a few chunks of time rather than consistently throughout the day) than group 2, as 50% (5/10) of physicians in group 1 batched their inbox work compared with 6% (1/17) in group 2 (X^2_1=4.03; P=.045). The rate of switching windows within the EHR was not statistically different among the 3 groups (P=.24), with all groups switching windows 4-4.5 times per minute of EHR use, on average. The groups spent different amounts of time per message (P=.004). The time per message was higher for group 1 (mean 0.46 min, SD 0.11 min) than for group 2 (mean 0.35 min, SD 0.06 min) and group 3 (mean 0.38 min, SD 0.07 min). Groups 2 and 3 did not differ significantly (P=.21). In terms of inbox message types, there were statistically significant differences among groups in patient-initiated messages (P=.02), with group 3 spending a higher average percentage of their inbox time on patient-initiated messages than group 1, and no differences for other group pairs (Table 1).

Stress Patterns

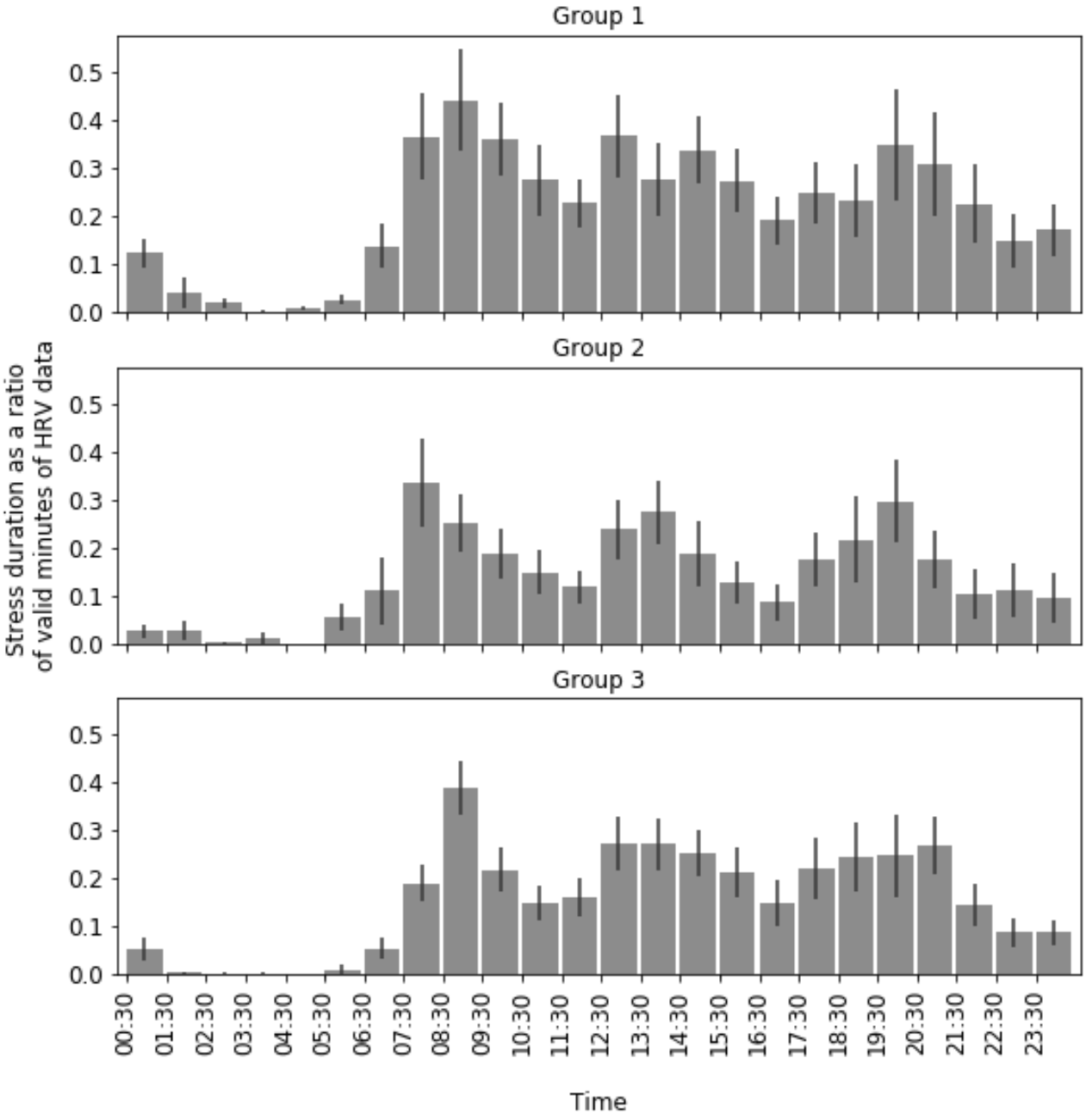
Visualizing stress patterns throughout the day showed that stress was high at the beginning of the workday. The first hour of work (8:30 AM to 9:30 AM) had an average stress duration of 35% of the hour (SD 26%; SE 4%). Stress then started to decrease until the lunch hour and increased again at the start of the afternoon clinic shift. Toward the end of the workday, the stress duration decreased. There was another increase in stress in the evening, followed by a decrease in stress at night and during typical sleep hours (Figure 2). This *3-wave* pattern of daily stress was consistent across the 3 work patterns, although group 2 had their highest stress an hour earlier (ie, 7:30 AM to 8:30 AM) than the other groups (Figure 2).

There was a difference in the average duration of stress during work hours among the groups (Kruskal-Wallis; P=.02). A posthoc comparison showed that group 1, the group with the highest after-hours inbox work duration, had a longer duration of stress during work hours than group 2 and group 3, with 33% (SD 27%) of work hours for group 1 being stressful (80 out of 243 min of valid HRV data indicated medium to high stress) compared with the 18% (SD 18%) for group 2 (47 out of 265 min of valid HRV data) and 22% (SD 24%) for group 3 (58 out of 265 min of valid HRV data). There was no significant difference between group 2 and 3 (P=.73). The number of valid minutes of HRV measurements was not significantly different across groups.

On average, physicians missed 45% (SD 20%; 9.4 out of 21) of the experience sampling prompts over the study period. Of the 485 submitted responses, 188 (38.8%) reported a stress level of over 50% (the midpoint of the slider). There was no significant difference in the average daily self-reported stress across the 3 inbox work patterns (P=.99).

Finally, in the survey on inbox management strategies and stress, physicians reported that 60% (SD 19%) of their work-related distress came from inbox management. Regarding the question of how distressful they find inbox management overall, of the 44 physicians, 19 (43%) said it was moderately stressful, 15 (34%) said it was very stressful, 6 (14%) said it was extremely stressful, and 4 (9%) said it was not very stressful. There were no statistically significant differences in survey responses across the 3 inbox work patterns.

Figure 2. Workday stress patterns of each group. Error bars represent the SE of the mean. HRV: heart rate variability.



EHR Use Characteristics Associated With Stress

We investigated detailed EHR use characteristics associated with stress using a mixed effects model, with workdays as the unit of analysis. The model showed that fixed effects accounted for 15% of the variation in duration of stress during work hours (Table 2). The physician's age, sex, and FTE worked were not associated with stress. The rate of switching windows when using the EHR was positively associated with stress (P=.001). Time spent on inbox work during work hours was positively associated with stress (P<.001), whereas time spent on other EHR activities during work hours was negatively (but very weakly) associated with stress (P<.001). Inbox work outside of work hours was positively associated with stress during work hours (P<.001). Interestingly, the proportion of inbox time spent on patient messages was not associated with stress. Surprisingly, batching inbox work for the day was also positively associated with stress (P<.001). Finally, days of the week were predictive of stress, with Mondays and Thursdays negatively associated with stress, whereas Tuesdays and Wednesdays positively associated with stress (P<.001 for each).

Table 2. Generalized linear mixed effects regression model.

Fixed effects[a]	β (SE)	Standard β[b]	*P* value
Full-time equivalent	1.94 (1.39)	.27	.16
Age	−.01 (.02)	−.05	.79
Female	.45 (.38)	.21	.24
Window switching rate	.1 (.03)	.08	.001
Work hours inbox duration	.003 (.001)	.08	<.001
Work hours noninbox EHR[c] duration	−.002 (0)	−.06	<.001
Nonwork hours inbox duration proportion	.35 (.07)	.09	<.001
Patient messages proportion	−.09 (.08)	−.01	.28
Batching	.13 (.03)	.06	<.001
Monday	−.22 (.04)	−.10	<.001
Tuesday	.16 (.03)	.06	<.001
Wednesday	.53 (.03)	.20	<.001
Thursday	−.13 (.04)	−.05	<.001

[a]The dependent variable is duration of stress during work hours. Friday is the reference category for the variable *day of week*.

[b]Standard β is the standardized coefficient.

[c]EHR: electronic health record.

Discussion

Principal Findings

To our knowledge, this study is the first to measure physician stress using wearable sensors over several days of outpatient practice and the first to identify distinct EHR inbox work patterns and their associations with stress. Although the topic of EHR use and stress (specifically, self-reported burden, burnout, workload, and well-being) has been addressed in previous studies, this study is novel in that we measured stress unobtrusively and continuously through physiologic measures and used system logs to gain detailed insight about EHR use factors associated with stress. Higher rates of EHR window switching, longer inbox work duration, and a higher proportion of inbox work done outside of work hours were associated with higher stress. Daily stress patterns showed 3 waves of stress: in the first hour of work, at or after lunch hours, and in the evening.

In addition, we found that physicians fell into 3 groups with different patterns of inbox work. Some physicians tended to do most of their inbox work within work hours, whereas others did inbox work before or after but contiguous to work hours. The third group did inbox work in late evenings. These groups differed in characteristics such as inbox work batching, time per message, and the proportion of inbox time spent on patient messages. Physicians who did most of their inbox work outside of work hours were more likely to batch email and spend more time per message, whereas physicians who mostly do their inbox work within work hours were more likely to continually check their inbox throughout the workday, potentially in the short periods of time between patient appointments, and spent less time per message. The group that did most of their inbox work outside of work hours had the longest stress duration during work hours.

A strength of this study is that we measured stress using 3 different methods. The HRV-based stress provided a continuous timestamped stress measure that could be correlated with inbox use patterns throughout the day, the experience sampling measure provided momentary self-assessment of stress 3 times a day, and the survey provided a reflective measure on perceived overall stress related to inbox work. HRV-based stress differed across groups but self-report measures did not. It is well established in the literature that short-term self-reported (ie, perceived) stress and acute physiological stress do not always align linearly in daily life settings [48-50]; however, both are important to monitor as they both have health and well-being implications [51-54].

Comparison With Previous Work

Previous studies on EHR use patterns have quantified the time spent on different EHR activities within and outside of work hours [14,24]. However, variation among physicians is not well studied, and no previous study has attempted to characterize physicians based on their patterns of daily inbox use. One study [16] found that physician-to-physician variation explains most of the variability in EHR use time. We extend the findings on the variation in EHR use, focusing on inbox use and comparing physician characteristics across work patterns based on work hours and after-hours EHR inbox use. Aligned with previous findings [16], we did not find differences in physicians' sex distributions between the group with the longest after-hours inbox time and the group with the shortest after-hours inbox time. We also did not find differences based on FTE, contrary to previous findings [16] that more work relative value units generated by physicians (another measure of workload) were associated with more EHR time after work hours.

Most studies use basic measures to characterize EHR usage, such as the duration of time [14,15,55]. In one study, researchers used more complex measures to characterize mobile EHR usage, such as the number of log-ins and features used and usage paths (ie, the frequency and complexity of consecutive actions) [56]. They compared doctors across medical specialties and found that physicians other than surgeons had more diverse mobile EHR usage patterns with higher complexity and repetitive loops compared to surgeons [56]. In this study, we also used detailed EHR and inbox usage characteristics such as window switching, inbox work batching, the time per message, message types, and the time distribution between work and nonwork hours. Our finding that the window switching rate was positively associated with stress could reflect the complexity and repetitiveness of physicians' EHR interactions, as indicated in prior work [56], and the efficiency issues often associated with physicians' satisfaction with EHRs [57]. Another study on EHR inbox burden [8] also reported that excessive steps were needed to process messages and that physicians recommended reducing the number of mouse clicks necessary to process messages.

A recent study suggested a relationship between patient call messages and clinician burnout [58]. Their category of patient messages included all messages related to patient care tasks, such as phone calls, refill requests, and patient care forms. In our study, the category of patient messages included only patient-initiated messages and was not found to be associated with stress, although it comprised most of the inbox time for physicians.

It is not surprising that the differences among groups in HRV-based stress did not align with self-reported perceived stress. Previous studies have noted several issues in the interrelationship between perceived and physiological stress [59]. For example, the timing of the perceived stress prompt (before, during, or after a stressor event) could determine whether and how perceived stress correlates with physiological stress measured during the stressor event [60-62]. This has important implications for real-time stress monitoring for physicians, as it suggests that daily prompts to measure perceived stress in situ could fail to capture physiological stress. Increased and prolonged physiological stress reactions are associated with several health and well-being risks [63].

The results also suggest practical implications for organizational changes and system design. Previous studies have recommended a fundamental redesign of the EHR to improve data entry and retrieval [11]. On the basis of our finding that window switching is associated with stress, a redesign that minimizes the need to navigate to different windows to record or obtain information may be beneficial. For example, contextual information for inbox messages can be made visible from the inbox [8]. Our findings lend support to recommendations from a previous study to automate frequently performed actions such as message routing and leverage team support for inbox management [8]. Allocating time for inbox management within work hours, also recommended in a previous study, may also help reduce stress [8].

Limitations

In this study, the regression model with EHR use characteristics explained 15% of the variation in duration of stress during work hours, which is a considerable proportion given the myriad factors that can potentially influence stress. However, stress was likely to have also been influenced by other variables that were beyond the scope of this study. In addition, the associations we observed between stress and window switching, inbox work duration, and inbox work outside work hours do not necessarily prove that the latter factors cause stress. It is possible that physicians who are busier during work hours have more stress and also make more window switches, have more inbox work, and have to do more inbox work outside work hours.

HRV-based measures are affected by several factors, such as health and physical activities. Although we tried to control these effects with our participant inclusion criteria and by removing periods that had physical activity registered by the wearable device, it is possible that carry-over effects of physical activity are still present in the HRV data of sedentary moments. Moreover, removing periods with physical activities could have removed periods when psychological stress was experienced. For example, walking to an important meeting could be mentally stressful but it will not be captured in our data because of the elimination of periods when walking is detected.

HRV data were excluded during periods of physical activities and were occasionally missing because of sensors losing contact with the skin. We set a minimum threshold (measurement period) of 20 minutes of valid data per hour for hourly stress measures and 2 hours for daily stress measures. Although not complete, we do feel that this is a reasonable proxy for the stress experience of that hour and day and a reasonable mitigation method for missing data.

Inbox use patterns might differ from one setting to another based on the organization's policies and norms. For example, the medical group where this study was conducted encouraged patients to use EHR portal messages to communicate with physicians. Simultaneously, system-generated messages and administrative reminders are kept to a minimum whenever possible. Thus, the distribution of different message types may differ from that in other settings. These factors must be considered when generalizing our findings.

Finally, some physicians might have had panel management time (ie, time designated by departments specifically for tasks such as inbox management) incorporated within their work hours. In this study, we did not have access to data on panel management time. Thus, we cannot make assumptions about why inbox work patterns differed among physicians. We can only report the relationship of these different work patterns with stress.

Conclusions

This study is the first to use continuous and unobtrusive measures of stress to evaluate associations between EHR inbox use and stress among physicians. A total of 3 potentially modifiable factors were associated with stress: window switching, inbox work duration, and inbox work outside work hours. These findings have implications for research and

organizational policies on stress measurement and EHR inbox management time and EHR system design.

Acknowledgments

The authors would like to thank MegAnn McGinnis, MPH, for assisting with participant enrollment. The authors would also like to thank Yi-Fen Irene Chen, MD; Sameer Awsare, MD; and Brian Hoberman, MD, of The Permanente Medical Group (TPMG) for their sponsorship and support of this work. This work was supported by The Permanente Medical Group via its Delivery Science Grants Program and by the National Science Foundation under grant 1704889. The funder (TPMG) supported the effort of the study team and had no specific requirements regarding the interpretation of results or framing of the manuscript. The National Science Foundation grant partially supported FA in the collection, management, analysis, and interpretation of data and GM in interpretation, preparation, and review. The views, opinions, and findings contained in this publication are those of the authors and do not necessarily reflect the views of Kaiser Permanente and should not be construed as an official position, policy, or decision of Kaiser Permanente unless designated by other documentation. No official endorsement was provided.

Authors' Contributions

All authors contributed to conceptualizing the study, interpreting the results, and critically editing the manuscript. FA contributed to designing the study, set up the technical infrastructure, analyzed the data, and wrote the manuscript. SP, MFM, and JAE contributed to the design of the study, facilitated data collection, and participated in the interpretation of results. EMW collected and preprocessed EHR log data. TAL, MER, and GM obtained funding and supervised the study.

Conflicts of Interest

None declared.

References

1. Murphy DR, Meyer AN, Russo E, Sittig DF, Wei L, Singh H. The burden of inbox notifications in commercial electronic health records. JAMA internal medicine 2016;176(4):560.
2. Renaud K, Ramsay J, Hair M. "You've got e-mail!" ... shall I deal with it now? Electronic mail from the recipient's perspective. International Journal of Human-Computer Interaction 2006;21(3):332.
3. Barley SR, Meyerson DE, Grodal S. E-mail as a source and symbol of stress. Organization Science 2011;22(4):906.
4. Mark G, Voida S, Cardello A. "A pace not dictated by electrons": an empirical study of work without email. In: CHI '12: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2012 May Presented at: CHI '12: SIGCHI Conference on Human Factors in Computing Systems; 05/2012; Austin, Texas, USA p. 555-564. [doi: 10.1145/2207676.2207754]
5. Tai-Seale M, Dillon EC, Yang Y, Nordgren R, Steinberg RL, Nauenberg T, et al. others. Physicians? Well-Being Linked To In-Basket Messages Generated By Algorithms In Electronic Health Records. Health Affairs ? 2019;38(7):1078.
6. Lieu TA, Freed GL. Unbounded?Parent-Physician Communication in the Era of Portal Messaging. JAMA pediatrics ? 2019;173(9):812.
7. Gregory ME, Russo E, Singh H. Electronic Health Record Alert-Related Workload as a Predictor of Burnout in Primary Care Providers. Appl Clin Inform 2017 Jul 05;8(3):686-697 [FREE Full text] [doi: 10.4338/ACI-2017-01-RA-0003] [Medline: 28678892]
8. Murphy DR, Satterly T, Giardina TD, Sittig DF, Singh H. Practicing Clinicians' Recommendations to Reduce Burden from the Electronic Health Record Inbox: a Mixed-Methods Study. J Gen Intern Med 2019 Sep;34(9):1825-1832 [FREE Full text] [doi: 10.1007/s11606-019-05112-5] [Medline: 31292905]
9. Gross JJ, John OP. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. J Pers Soc Psychol 2003 Aug;85(2):348-362. [doi: 10.1037/0022-3514.85.2.348] [Medline: 12916575]
10. Sato H, Kawahara JI. Selective bias in retrospective self-reports of negative mood states. Anxiety Stress Coping 2011 Jul;24(4):359-367. [doi: 10.1080/10615806.2010.543132] [Medline: 21253957]
11. Colicchio TK, Cimino JJ, Del Fiol G. Unintended Consequences of Nationwide Electronic Health Record Adoption: Challenges and Opportunities in the Post-Meaningful Use Era. J Med Internet Res 2019 Jun 03;21(6):e13313 [FREE Full text] [doi: 10.2196/13313] [Medline: 31162125]
12. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction. Presented at: Mayo Clinic Proceedings Elsevier; .?848; 2016 p. 836.
13. Gardner R, Cooper E, Haskell J, Harris D, Poplau S, Kroth P, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]

14. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan WJ, Sinsky CA, et al. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. The Annals of Family Medicine ? 2017;15(5):426.
15. Saag HS, Shah K, Jones SA, Testa PA, Horwitz LI. Pajama Time: Working After Work in the Electronic Health Record. Journal of general internal medicine ? 2019;1:2.
16. Attipoe S, Huang Y, Schweikhart S, Rust S, Hoffman J, Lin S. Factors Associated With Electronic Health Record Usage Among Primary Care Physicians After Hours: Retrospective Cohort Study. JMIR Hum Factors 2019 Sep 30;6(3):e13779. [doi: 10.2196/13779]
17. Adler-Milstein J, Zhao W, Willard-Grace R, Knox M, Grumbach K. Electronic health records and burnout: Time spent on the electronic health record after hours and message volume associated with exhaustion but not with cynicism among primary care clinicians. J Am Med Inform Assoc 2020 Apr 01;27(4):531-538 [FREE Full text] [doi: 10.1093/jamia/ocz220] [Medline: 32016375]
18. Heponiemi T, Kujala S, Vainiomäki S, Vehko T, Lääveri T, Vänskä J, et al. Usability Factors Associated With Physicians' Distress and Information System–Related Stress: Cross-Sectional Survey. JMIR Med Inform 2019 Nov 5;7(4):e13466. [doi: 10.2196/13466]
19. Vainiomäki S, Aalto A, Lääveri T, Sinervo T, Elovainio M, Mäntyselkä P, et al. Better Usability and Technical Stability Could Lead to Better Work-Related Well-Being among Physicians. Appl Clin Inform 2017 Dec 14;08(04):1057-1067. [doi: 10.4338/ACI-2017-06-RA-0094]
20. Khairat S, Burke G, Archambault H, Schwartz T, Larson J, Ratwani R. Focus Section on Health IT Usability: Perceived Burden of EHRs on Physicians at Different Stages of Their Career. Appl Clin Inform 2018 Apr;? ;09(02):347. [doi: 10.1055/s-0038-1648222]
21. Palojoki S, Pajunen T, Saranto K, Lehtonen L. Electronic Health Record-Related Safety Concerns: A Cross-Sectional Survey of Electronic Health Record Users. JMIR Med Inform 2016 May 06;4(2):e13 [FREE Full text] [doi: 10.2196/medinform.5238] [Medline: 27154599]
22. Chen Y, Ngo V, Harrison S, Duong V. Unpacking Exam-Room Computing: Negotiating Computer-Use in Patient-Physician Interactions. In: CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery; 2011 May Presented at: CHI '11: SIGCHI Conference on Human Factors in Computing Systems; 05/2011; Vancouver, BC, Canada p. 3343-3352. [doi: 10.1145/1978942.1979438]
23. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Smith MW, Murphy DR, et al. Primary care practitioners' views on test result management in EHR-enabled health systems: a national survey. J Am Med Inform Assoc 2013 Jul 01;20(4):727-735. [doi: 10.1136/amiajnl-2012-001267]
24. Akbar F, Mark G, Warton E, Reed M, Prausnitz S, East J, et al. Physicians' electronic inbox work patterns and factors associated with high inbox work duration. J Am Med Inform Assoc 2020 Oct 15:229. [doi: 10.1093/jamia/ocaa229] [Medline: 33063087]
25. Singh H, Spitzmueller C, Petersen NJ, Sawhney MK, Sittig DF. Information overload and missed test results in electronic health record-based settings. JAMA Intern Med 2013 Apr 22;173(8):702-704 [FREE Full text] [doi: 10.1001/2013.jamainternmed.61] [Medline: 23460235]
26. Murphy DR, Reis B, Sittig DF, Singh H. Notifications received by primary care practitioners in electronic health records: a taxonomy and time analysis. Am J Med 2012 Feb;125(2):209.e1-209.e7. [doi: 10.1016/j.amjmed.2011.07.029] [Medline: 22269625]
27. Lieu TA, Warton EM, East JA, Moeller MF, Prausnitz S, Ballesca M, et al. Evaluation of Attention Switching and Duration of Electronic Inbox Work Among Primary Care Physicians. JAMA Netw Open 2021 Jan 04;4(1):e2031856 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.31856] [Medline: 33475754]
28. Ray RD, McRae K, Ochsner KN, Gross JJ. Cognitive reappraisal of negative affect: converging evidence from EMG and self-report. Emotion 2010 Aug;10(4):587-592 [FREE Full text] [doi: 10.1037/a0019015] [Medline: 20677875]
29. Stone AA, Turkkan JS, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS. The science of self-report: Implications for research and practice. Mahwah, NJ: Lawrence Erlbaum Associates Publishers; 1999.
30. Rajendra Acharya U, Paul Joseph K, Kannathal N, Lim CM, Suri JS. Heart rate variability: a review. Med Biol Eng Comput 2006 Dec 17;44(12):1031-1051. [doi: 10.1007/s11517-006-0119-0] [Medline: 17111118]
31. Malik M, Bigger JT, Camm AJ, Kleiger RE, Malliani A, Moss AJ, et al. Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. European Heart Journal 1996 Mar 01;17(3):354-381. [doi: 10.1093/oxfordjournals.eurheartj.a014868]
32. Stys A, Stys T. Current clinical applications of heart rate variability. Clin Cardiol 1998 Oct;21(10):719-724 [FREE Full text] [doi: 10.1002/clc.4960211005] [Medline: 9789691]
33. Picard RW, Fedor S, Ayzenberg Y. Multiple Arousal Theory and Daily-Life Electrodermal Activity Asymmetry. Emotion Review 2015 Mar 02;8(1):62-75. [doi: 10.1177/1754073914565517]
34. Okkonen J, Heimonen T, Savolainen R, Turunen M. Assessing Information Ergonomics in Work by Logging and Heart Rate Variability. USA: Springer; 2018 Presented at: International Conference on Applied Human Factors and Ergonomics; 2017; Los Angeles, USA p. 425-436. [doi: 10.1007/978331960492341]

35. Mark G, Iqbal S, Czerwinski M, Johns P, Sano A, Lutchyn Y. Email Duration, Batching and Self-interruption: Patterns of Email Use on Productivity and Stress. In: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2016 Presented at: 2016 CHI Conference on Human Factors in Computing Systems; 2016; San Jose, California, USA p. 1717-1728. [doi: 10.1145/2858036.2858262]
36. Koldijk S, Sappelli M, Neerincx M, Kraaij W. Unobtrusive Monitoring of Knowledge Workers for Stress Self-regulation. In: Carberry S., Weibelzahl S., Micarelli A., Semeraro G. (eds) User Modeling, Adaptation, and Personalization. UMAP 2013. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2013.
37. Lyu Y, Luo X, Zhou J, Yu C, Miao C, Wang T, et al. Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. In: CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York, NY, USA: ACM Press; 2015 Apr Presented at: ACM Conference on Human Factors in Computing Systems; 2015; Seoul, Republic of Korea p. 857-866. [doi: 10.1145/2702123.2702399]
38. Mattingly S, Gregg J, Audia P, Bayraktaroglu A, Campbell A, Chawla N, et al. The Tesserae Project: Large-Scale, Longitudinal, In Situ, Multimodal Sensing of Information Workers. In: CHI'19 Extended Abstracts: Proceedings of CHI Conference on Human Factors in Computing Systems Extended Abstracts. New York, NY, USA: ACM; 2019 Presented at: CHI Conference on Human Factors in Computing Systems; 05/2019; Glasgow, Scotland UK. [doi: 10.1145/3290607.3299041]
39. Jessup G, Bian S, Chen Y-W, Bundy A. PIEL Survey application. 2012. URL: https://ses.library.usyd.edu.au/handle/2123/9490 [accessed 2020-07-07]
40. Stress and Recovery Analysis Method Based on 24-hour Heart Rate Variability Internet. Firstbeat Technologies Ltd. URL: https://assets.firstbeat.com/firstbeat/uploads/2015/10/Stress-and-recovery_white-paper_20145.pdf [accessed 2020-07-06]
41. Fuller D, Colwell E, Low J, Orychock K, Tobin MA, Simango B, et al. Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review. JMIR Mhealth Uhealth 2020 Sep 08;8(9):e18694 [FREE Full text] [doi: 10.2196/18694] [Medline: 32897239]
42. Gillinov S, Etiwy M, Wang R, Blackburn G, Phelan D, Gillinov A, et al. Variable Accuracy of Wearable Heart Rate Monitors during Aerobic Exercise. Med Sci Sports Exerc 2017 Aug;49(8):1697-1703. [doi: 10.1249/MSS.0000000000001284] [Medline: 28709155]
43. Claes J, Buys R, Avila A, Finlay D, Kennedy A, Guldenring D, et al. Validity of heart rate measurements by the Garmin Forerunner 225 at different walking intensities. J Med Eng Technol 2017 Aug;41(6):480-485. [doi: 10.1080/03091902.2017.1333166] [Medline: 28675070]
44. Dooley EE, Golaszewski NM, Bartholomew JB. Estimating Accuracy at Exercise Intensities: A Comparative Study of Self-Monitoring Heart Rate and Physical Activity Wearable Devices. JMIR Mhealth Uhealth 2017 Mar 16;5(3):e34 [FREE Full text] [doi: 10.2196/mhealth.7043] [Medline: 28302596]
45. Blascovich J, Tomaka J. The biopsychosocial model of arousal regulation. Advances in experimental social psychology 1996;28:1-51. [doi: 10.1016/s0065-2601(08)60235-x]
46. Reynolds D. Gaussian Mixture Models. In: Li S.Z., Jain A.K. (eds) Encyclopedia of Biometrics. Boston, MA: Springer, , MA; 2015.
47. Starczewski A, Krzyżak A. Performance Evaluation of the Silhouette Index. In: Artificial Intelligence and Soft Computing. ICAISC 2015. Lecture Notes in Computer Science, vol 9120. Cham: Springer International Publishing; 2015.
48. Kageyama T, Nishikido N, Kobayashi T, Kurokawa Y, Kaneko T, Kabuto M. Self-reported sleep quality, job stress, and daytime autonomic activities assessed in terms of short-term heart rate variability among male white-collar workers. Ind Health 1998 Jul;36(3):263-272 [FREE Full text] [doi: 10.2486/indhealth.36.263] [Medline: 9701906]
49. Hernandez RJ. Towards wearable stress measurement. PhD Dissertation. 2015. URL: https://dspace.mit.edu/bitstream/handle/1721.1/101849/942938376-MIT.pdf?sequence=1
50. Muaremi A, Arnrich B, Tröster G. Towards measuring stress with smartphones and wearable devices during workday and sleep. BioNanoScience ? 2013;3(2):183.
51. Rod N, Grønbaek M, Schnohr P, Prescott E, Kristensen T. Perceived stress as a risk factor for changes in health behaviour and cardiac risk profile: a longitudinal study. J Intern Med 2009 Nov;266(5):467-475 [FREE Full text] [doi: 10.1111/j.1365-2796.2009.02124.x] [Medline: 19570055]
52. Lagraauw HM, Kuiper J, Bot I. Acute and chronic psychological stress as risk factors for cardiovascular disease: Insights gained from epidemiological, clinical and experimental studies. Brain Behav Immun 2015 Nov;50:18-30. [doi: 10.1016/j.bbi.2015.08.007] [Medline: 26256574]
53. VanItallie TB. Stress: A risk factor for serious illness. Metabolism 2002 Jun;? ;51(6):45. [doi: 10.1053/meta.2002.33191]
54. Vogelzangs N, Beekman ATF, Milaneschi Y, Bandinelli S, Ferrucci L, Penninx BWJH. Urinary cortisol and six-year risk of all-cause and cardiovascular mortality. J Clin Endocrinol Metab 2010 Nov;95(11):4959-4964 [FREE Full text] [doi: 10.1210/jc.2010-0192] [Medline: 20739384]
55. Anderson J, Leubner J, Brown SR. EHR Overtime: An Analysis of Time Spent After Hours by Family Physicians. Fam Med 2020 Feb;52(2):135-137. [doi: 10.22454/FamMed.2020.942762] [Medline: 32050270]

56. Soh JY, Jung SH, Cha WC, Kang M, Chang DK, Jung J, et al. Variability in Doctors' Usage Paths of Mobile Electronic Health Records Across Specialties: Comprehensive Analysis of Log Data. JMIR Mhealth Uhealth 2019 Jan 17;7(1):e12041 [FREE Full text] [doi: 10.2196/12041] [Medline: 30664473]
57. Williams DC, Warren RW, Ebeling M, Andrews AL, Teufel Ii RJ. Physician Use of Electronic Health Records: Survey Study Assessing Factors Associated With Provider Reported Satisfaction and Perceived Patient Impact. JMIR Med Inform 2019 Apr 04;7(2):e10949 [FREE Full text] [doi: 10.2196/10949] [Medline: 30946023]
58. Hilliard R, Haskell J, Gardner R. Are specific elements of electronic health record use associated with clinician burnout more than others? J Am Med Inform Assoc 2020 Jul 01;27(9):1401-1410. [doi: 10.1093/jamia/ocaa092] [Medline: 32719859]
59. Hjortskov N, Garde AH, Ørbæk P, Hansen ?. Evaluation of salivary cortisol as a biomarker of self-reported mental stress in field studies. Stress and Health 2004 Apr 07;20(2):91-98. [doi: 10.1002/smi.1000]
60. Schlotz W, Kumsta R, Layes I, Entringer S, Jones A, Wüst S. Covariance between psychological and endocrine responses to pharmacological challenge and psychosocial stress: a question of timing. Psychosom Med 2008 Sep;70(7):787-796. [doi: 10.1097/PSY.0b013e3181810658] [Medline: 18725434]
61. Gaab J, Rohleder N, Nater U, Ehlert U. Psychological determinants of the cortisol stress response: the role of anticipatory cognitive appraisal. Psychoneuroendocrinology 2005 Jul;30(6):599-610. [doi: 10.1016/j.psyneuen.2005.02.001] [Medline: 15808930]
62. Oldehinkel A, Ormel J, Bosch N, Bouma E, Van Roon AM, Rosmalen J, et al. Stressed out? Associations between perceived and physiological stress responses in adolescents: the TRAILS study. Psychophysiology 2011 Apr;48(4):441-452. [doi: 10.1111/j.1469-8986.2010.01118.x] [Medline: 21361964]
63. Logan JG, Barksdale DJ. Allostasis and allostatic load: expanding the discourse on stress and cardiovascular disease. J Clin Nurs 2008 Apr;17(7B):201-208. [doi: 10.1111/j.1365-2702.2008.02347.x] [Medline: 18578796]

Abbreviations

EDA: electrodermal activity
EHR: electronic health record
FTE: full-time equivalent
HRV: heart rate variability
PCP: primary care physician
TPMG: The Permanente Medical Group

Edited by C Lovis; submitted 01.09.20; peer-reviewed by M Murero, N Fijacko, C Wang; comments to author 23.12.20; revised version received 02.02.21; accepted 21.03.21; published 28.04.21.

Please cite as:
Akbar F, Mark G, Prausnitz S, Warton EM, East JA, Moeller MF, Reed ME, Lieu TA
Physician Stress During Electronic Health Record Inbox Work: In Situ Measurement With Wearable Sensors
JMIR Med Inform 2021;9(4):e24014
URL: https://medinform.jmir.org/2021/4/e24014
doi:10.2196/24014
PMID:33908888

Original Paper

Weight-Based Framework for Predictive Modeling of Multiple Databases With Noniterative Communication Without Data Sharing: Privacy-Protecting Analytic Method for Multi-Institutional Studies

Ji Ae Park[1], MSc; Min Dong Sung[1], MD; Ho Heon Kim[1], BSc; Yu Rang Park[1], PhD

Department of Biomedical System Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Yu Rang Park, PhD
Department of Biomedical System Informatics
Yonsei University College of Medicine
50-1 Yonsei-ro, Seodaemun-gu
Seoul
Republic of Korea
Phone: 82 2 2228 2493
Email: yurangpark@yuhs.ac

Abstract

Background: Securing the representativeness of study populations is crucial in biomedical research to ensure high generalizability. In this regard, using multi-institutional data have advantages in medicine. However, combining data physically is difficult as the confidential nature of biomedical data causes privacy issues. Therefore, a methodological approach is necessary when using multi-institution medical data for research to develop a model without sharing data between institutions.

Objective: This study aims to develop a weight-based integrated predictive model of multi-institutional data, which does not require iterative communication between institutions, to improve average predictive performance by increasing the generalizability of the model under privacy-preserving conditions without sharing patient-level data.

Methods: The weight-based integrated model generates a weight for each institutional model and builds an integrated model for multi-institutional data based on these weights. We performed 3 simulations to show the weight characteristics and to determine the number of repetitions of the weight required to obtain stable values. We also conducted an experiment using real multi-institutional data to verify the developed weight-based integrated model. We selected 10 hospitals (2845 intensive care unit [ICU] stays in total) from the electronic intensive care unit Collaborative Research Database to predict ICU mortality with 11 features. To evaluate the validity of our model, compared with a centralized model, which was developed by combining all the data of 10 hospitals, we used proportional overlap (ie, 0.5 or less indicates a significant difference at a level of .05; and 2 indicates 2 CIs overlapping completely). Standard and firth logistic regression models were applied for the 2 simulations and the experiment.

Results: The results of these simulations indicate that the weight of each institution is determined by 2 factors (ie, the data size of each institution and how well each institutional model fits into the overall institutional data) and that repeatedly generating 200 weights is necessary per institution. In the experiment, the estimated area under the receiver operating characteristic curve (AUC) and 95% CIs were 81.36% (79.37%-83.36%) and 81.95% (80.03%-83.87%) in the centralized model and weight-based integrated model, respectively. The proportional overlap of the CIs for AUC in both the weight-based integrated model and the centralized model was approximately 1.70, and that of overlap of the 11 estimated odds ratios was over 1, except for 1 case.

Conclusions: In the experiment where real multi-institutional data were used, our model showed similar results to the centralized model without iterative communication between institutions. In addition, our weight-based integrated model provided a weighted average model by integrating 10 models overfitted or underfitted, compared with the centralized model. The proposed weight-based integrated model is expected to provide an efficient distributed research approach as it increases the generalizability of the model and does not require iterative communication.

(JMIR Med Inform 2021;9(4):e21043) doi:10.2196/21043

KEYWORDS

multi-institutional study; distributed data; data sharing; privacy-protecting methods

Introduction

Multi-institutional studies have many advantages in that they can increase the generalizability and reproducibility of clinical results. Studies based on geographically and demographically diverse populations using multi-institutional data are increasingly common and necessary to improve generalizability [1]. This increases the applicability of study results to other settings or with other samples, as sampling bias is reduced. Sampling bias occurs when patient and disease characteristics differ from the represented patient population, and it commonly occurs in electronic health record–derived databases from single institutions, as patient populations reflect the local socioeconomic environment or specialty interests of hospitals [2].

Data accumulated in multiple institutions should be shared to realize the potential of big data in medicine. Big biomedical data networks, such as the patient-centered Scalable National Network for Effectiveness Research clinical data research network [3], Scalable Architecture for Federated Translational Inquiries Network [4], and Electronic Medical Records and Genomics (eMERGE) network [5], have been established to enable cross-institutional biomedical studies [6]. As big data are relative to volume, variety, and velocity, their serviceability depends on combining and analyzing rapidly growing data sources stored in different places via these data networks.

However, the availability of such large volumes of data is associated with privacy issues. Privacy must be protected when sensitive biomedical data are being used for research purposes, and this requires implementing several safeguards [7]. To overcome the 2 conflicting problems of privacy and data usage, a methodological solution that can analyze all partitioned data without data sharing should be considered. The current approaches toward constructing models based on multi-institution data by solving the privacy concern on patient-level data distributed across institutions can be primarily categorized into distributed computing approaches, which require iterative communication between institutions, and approaches that do not require an iterative process in terms of communication efficiency.

Among the methods that use distributed computing, federated learning has recently been proposed as a promising solution. It is a distributed computing method wherein several clients collaboratively train a shared global model with the coordination of a central server [8]. A client can be a mobile or edge device, not an institution; however, if the client is a reliable institution, it is classified as cross-silo federated learning [9]. Cross-silo federated learning aims to solve an optimization problem by setting the objective function [10] for the centralized model. In general, this optimization problem can be managed by stochastic gradient descent. Each client computes the local gradient and returns it to the server for aggregation and, accordingly, the global parameter is updated [8]. This process is repeated until the parameter converges. Various studies have also developed algorithms to establish statistical models, such as GLORE (Grid Binary LOgistic Regression) [11] for logistic regression, grid multicategory response logistic models [12] for ordinal and multinomial logistic regressions, and WebDISCO (a web service for distributed Cox model learning) [13] for the Cox model. In these studies, the global likelihood function of the centralized model was divided into local likelihood functions for each institution; to estimate the parameter maximizing the global likelihood function, the nonsensitive intermediary results were iteratively exchanged between the central server and the institutions using the Newton–Raphson method [14]. These methods can guarantee the precision of the models; however, the solutions may leak patient information owing to the disclosure of the information matrix and score vectors during iterative model learning [6].

The noniterative approach aggregates the intermediate results required for building a global model without requiring an iterative process. A typical method is meta-analysis [15], which is a conventional statistical analysis. Meta-analysis is used to estimate the effect size (eg, correlation coefficient, odds ratio [OR], and hazard ratio) of the overall institution, rather than building a predictive model. The overall effect size is estimated by averaging the effect sizes that are estimated from each institution; this method has been widely used in various studies [16-19] based on the common data model adopted by the Observational Health Data Sciences and Informatics Consortium [20]. Further, by constructing a surrogate likelihood, ODAL (one-shot distributed algorithm to perform logistic regression) [21] and ODAC (one-shot distributed algorithm for Cox model) [22] have been proposed for the logistic and Cox models, respectively; these models can estimate the global parameters in a noniterative manner without using the Newton–Raphson method. By contrast, MCCG (the multicenter collaboration gateway) [23,24], which focuses on developing a prediction model, was proposed to improve the predictive performance of a specific target institution. Rather than constructing the centralized model, this algorithm proposed a method of aggregating the models of each institution such that they are trained in a single target institution to improve the predictive performance in that target institution.

In this study, we focus on developing a noniterative algorithm that can construct predictive models from different sources without sharing horizontally partitioned data, where patient-level data are divided for the same medical information. The proposed model, referred to as the weight-based integrated model, is a predictive model reflecting the characteristics of various populations in multiple institutions without compromising privacy. We evaluated the proposed weight-based integrated model based on 2 aspects: (1) To confirm whether it provides a weighted average model with all characteristics of multi-institutional data, we evaluated its similarity with the centralized model that was developed by combining all institutional data, compared with models from different institutions, in terms of the predictive power and parameter estimation. (2) To confirm whether the proposed weight-based

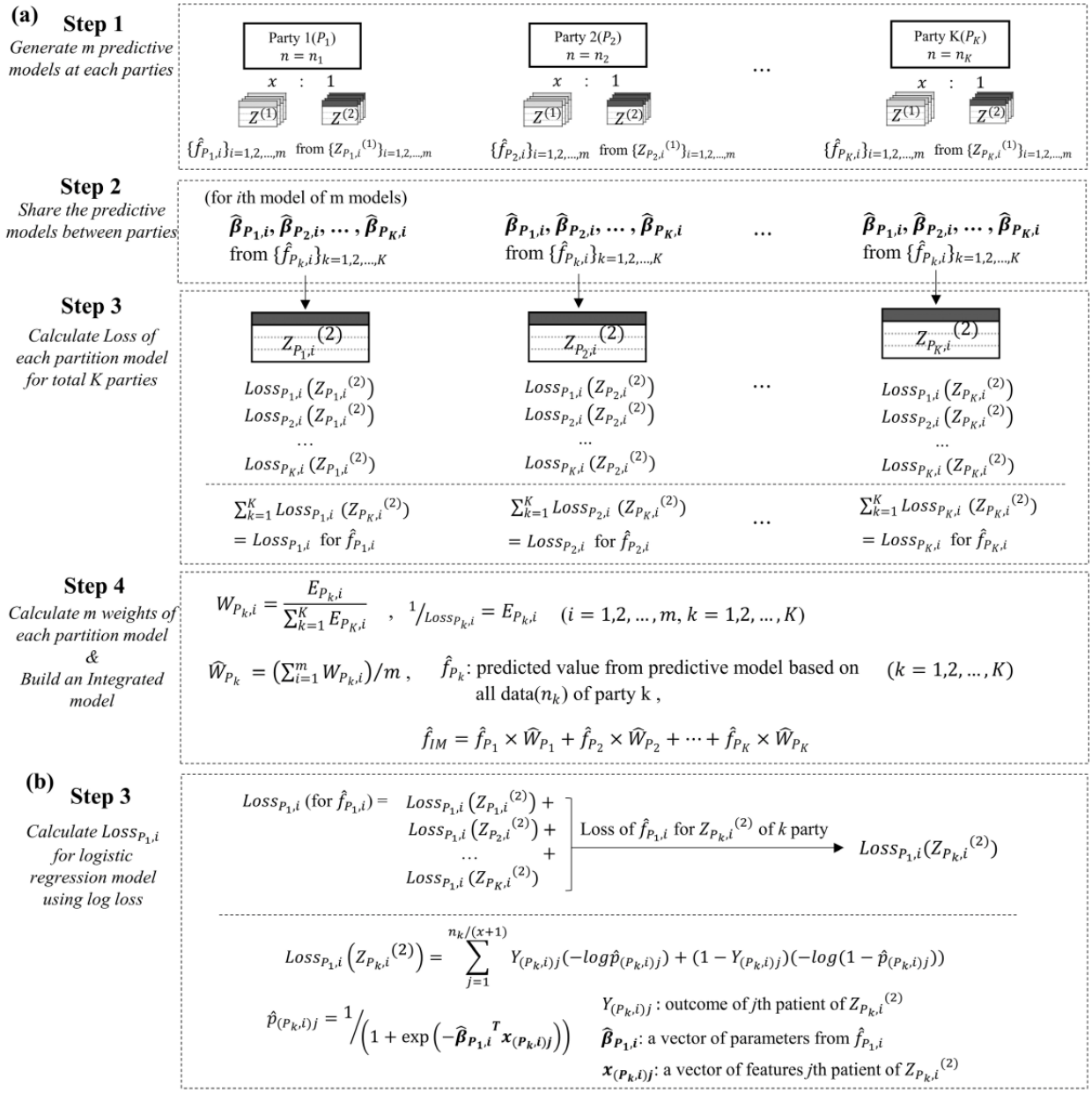
integrated model improves the average predictive performance by building a predictive model with generalizability, we compared the predictive power of the weight-based integrated model with that of the central model, as well as the models of each institution that were used to build weight-based integrated model, through external validation.

Methods

Weight-Based Integrated Model

The proposed weight-based integrated model involves a 4-step process (Figure 1). In step 1, 2 data sets are generated by each party to estimate a predictive model and to evaluate the performance. In step 2, the parameters estimated by each party are shared between the parties. In step 3, a loss value for the model of each party is calculated by fitting the model to the data set of the entire party. The larger the loss value from the model of each party, the smaller the weight of the model. In step 4, the weight-based integrated model is constructed based on the weight of each party. To describe the 4 steps in detail, assume K partitioned data, each of size n_k, and let P_k, $1 \le k \le K$, denote the kth partitioned data.

Figure 1. (A) Overall process of the weight-based integrated model. (B) Step 3 of the weight-based integrated model showing the process for calculating the weight using the log loss as a criterion to measure the model performance in the logistic regression model.



Step 1

Randomly split the *k*th party of size n_k into 2 parts—the first part is $Z^{(1)}$ with size $(n_k x)/(x+1)$, and the second part is $Z^{(2)}$ with size $(n_k)/(x+1)$. Here, $Z^{(1)}$ is used to estimate any predictive model f, whereas $Z^{(2)}$ is used to measure the predictive performance of the estimated model $\hat{f}$ obtained from $Z^{(1)}$. The data set ($Z^{(1)}$, $Z^{(2)}$) is generated m times for each P_k. Let i, $1 \leq i \leq m$, denote the number of data sets.  represents the *i*thdata set ($Z^{(1)}$, $Z^{(2)}$) of P_k.

Step 2

 is the *i*th model of P_k,estimated using  , and  is a vector of parameters estimated from  . The K parties share m vectors of parameters,  , with each other.

Step 3

In the *k*th party, fit the K models,  , including their model,  , which is estimated from  and sent from step 2 to the *i*th  . Subsequently, calculate the loss value for each of the K models.

 represents loss fitting  to  . Loss for total  of  is calculated as  and represents  . The loss function can vary depending on the model. For binary classification models (eg, the logistic regression model), the following log loss function [25], which is calculated as the negative log likelihood for probability predictions, can be used. The log loss function (or negative log likelihood function) of the logistic regression model for N patients is expressed as

where $p_i = 1/(1 + \exp[-\beta^T x_i])$ is the probability of outcome of interest, β^T is a vector of parameters, x_i is a vector of features of the *i*th patient, and $y_i$is a binary outcome of the *i*th patient. Figure 1B presents the process of calculating the loss for the *i*th model of party 1 (ie, ) using the log loss function.

To make the weight larger as the loss becomes smaller, we define  as the inverse of  , and  represents the goodness of fit for all K parties of the model of the corresponding weight.

Step 4

The  , represented by *i*th weight of the partition model of P_k for the integrated model, is calculated as follows:

where  represents the final weight of the partition model based on P_k, and can be obtained by averaging the  . The weight-based integrated model,  , is estimated as follows, using  , which represents a predicted value from the partition model of P_k based on the total $n_k$data. Note that  .

The weight calculated by the weight-based integrated model is determined by 2 factors: the data size of the party (ie, the ratio of data size to central data) and how well the model of the party fits into the data of the other parties (ie, the goodness of fit to all parties of the model from each party). In case of a party k with relatively large data, as the proportion of data of party k in the total  increases,  of the model of party k becomes small, and  becomes larger than the other parties. In other words, a party with a large data set has a large weight, and that with a small data set has a small weight. Further, the better the model of party k is fitted to the data of other parties, the smaller the loss values and the greater the weights. These characteristics of weights are demonstrated in the experiments based on simulations and real data.

The parameters of the model can be also estimated based on weights from the weight-based integrated model process. In step 3, the models and weights of K parties are generated for every i repetitions. Further, the weight-based parameter can be estimated based on the *i*th weights, , and *i*th vectors of parameters,  , estimated from each K party (I = 1, 2, ..., m). Let  be the *i*th vector of weight-based parameters. Then,  is calculated using  ; that is, parameter estimation in the weight-based integrated model is performed by calculating the weighted average of the parameters that is estimated by the models of each institution based on the weights on models of each institution. A point estimation and 95% CI estimation of a weight-based parameter can be performed using the average and (lower 2.5%, upper 97.5%) of m weight-based parameters, respectively.

Simulation Study

We performed 3 simulations. The first simulation aimed to validate the optimal number of repetitions of the weight. The second and third simulations were performed to show the features of the weight calculated using the weight-based integrated model and to compare with other weighting methods. For all simulations, the standard logistic regression model was used, and 5 features were set. Three features were sampled from binomial (1, 0.5), and 2 features were sampled from normal (0, 1). The outcome was generated from the binomial (1, p), where  , given 5 features (X) and 6 parameters (β). We set the 6 parameters to values from –2 to 2. The values of the parameters were set to adjust the homogeneous or heterogeneous characteristics between the parties.

In the first simulation, to set an optimal m associated with the number of repetitions of a weight per party, we examined the change in weight by adjusting the repetition m under each partitioned data size n for the following sizes: 200, 400, 600, 800, and 1000. A total of 23 scenarios were considered, with

the number of repetitions being 5 units from 5 to 50 and 50 units from 100 to 700. Three parties (A, B, and C) were considered. In this simulation, the adjustment of the homogeneous or heterogeneous characteristics of each party is not an important factor. Therefore, we generated 6 parameters for each party uniformly from [–2, 2].

The second simulation was performed to confirm the change pattern of the weights by adjusting 2 factors: the data size and the goodness of fit of the model from each party. In this simulation, we considered 2 scenarios. In the first scenario, we generated 3 parties (A, B, and C) with data sizes of 1000. One of the 3 parties was generated with a biased feature by adjusting the parameters for sampling. All 6 parameters of parties A and B were set the same. By setting 5 conditions of parameters, from parameter 1 to parameter 5, the biased degree of party C was increased as it was adjusted from parameter 1 to parameter 5. All 6 parameters of parties A and B were set equal to 1 at 5 conditions, and the parameters of party C were set to 1 at the condition of parameter 1, 0.5 at the condition of parameter 2, –0.5 at the condition of parameter 3, –1 at the condition of parameter 4, and –2 at the condition of parameter 5. That is, under the same data size, the change degree of the weights was confirmed by gradually deteriorating the goodness of fit for the entire data of the biased party C. In the second scenario, after setting one of the 3 parties to be biased, we changed the condition of data size to check the change degree of the weights according to the data size. The 6 parameters of parties A and B were set to 1, and all of party C were set to –2.

In the third simulation, we compared the weight of the weight-based integrated model with other comparable weighting methods to show the unique characteristics of the weight-based integrated model. This simulation aims to confirm to what extent the predictive performance of the integrated model using each weighting method is similar to that of the centralized model. We referred to an approach [26] of weighting strategies that investigated replicability of the performance of predictors across studies through ensembles of prediction models trained on different studies as the weights used in comparison. We chose 3 comparable weights in the approach [26] of weighting strategies: simple average (Avg), average weighted by study sample size (n-Avg), and average weighted by cross-study performance (CS-Avg). For K parties, with total data size N and kth party of size n_k, Avg assigns a weight of $1/K$ to each party, and n-Avg assigns a weight of n_k/N to each party. In addition, similar to the weight of the weight-based integrated model, CS-Avg constructs a predictive model for each party and then calculates the weight based on predictive performance for other parties. In calculating the performance of models for each party, the party used in the model is excluded. Further, the smaller the performance, the smaller the weight assigned, and the model with the lowest performance is assigned a weight of 0. An averaged value, such as the mean squared error, is used for performance measurement. For application to the logistic model of CS-Avg, we measured the performance by dividing the log loss function by the data size.

We performed 200 simulations under the same conditions. Four parties (A, B, C, and D) were constructed to build a predictive model, and another 4 validation parties were constructed to measure predictive performance. In addition, we assumed 2 scenarios, similar to the second simulation, to show the characteristics of each weight. While adjusting the data characteristics of parties under the same data sizes, and data sizes of parties under the same data characteristics, we observed the change patterns of weights and predictive performance of each weighting method. In the first scenario, the data sizes of the 4 parties were all set to 500. The 6 parameters, [β_0, β_1, β_2, β_3, β_4, β_5], of parties A and B were set to [0, 2, 2, 2, 2, 2], and the data characteristics of parties C and D were adjusted under the following 3 conditions: (1) 6 parameters—[0, 2, 2, 2, 2, 2], outcome generation: binomial (1, p); (2) 6 parameters—[0, –2, –2, 2, 2, –2], outcome generation: binomial (1, p); and (3) 6 parameters—[0, –2, –2, 2, 2, –2], outcome generation: binomial [1, min(0.5, p)]. The first condition, that is, (1), represents the same characteristics as parties A and B. By adjusting the parameter in (2) and the parameters and probability of generating an event in (3), the characteristics of parties C and D were gradually generated to be heterogeneous with parties A and B. In the second scenario, under the third condition of the first scenario, the data sizes of parties A and B were set to 500, and only the data sizes of parties C and D were changed to 500, 750, and 1000.

The data sizes of the 4 validation parties were all fixed at 500, and the data characteristics were the same as each condition of the first and the second scenarios. For example, the parameters of the 4 validation parties for condition (1) of the first scenario were set to [0, 2, 2, 2, 2, 2] in the same manner as parties A, B, C, and D. The average area under the receiver operating characteristic (ROC) curve (AUC) was measured for 4 validation parties to compare the similarity of the performance of each weighting method with that of the centralized model.

Experiment Using Real Horizontally Partitioned Data

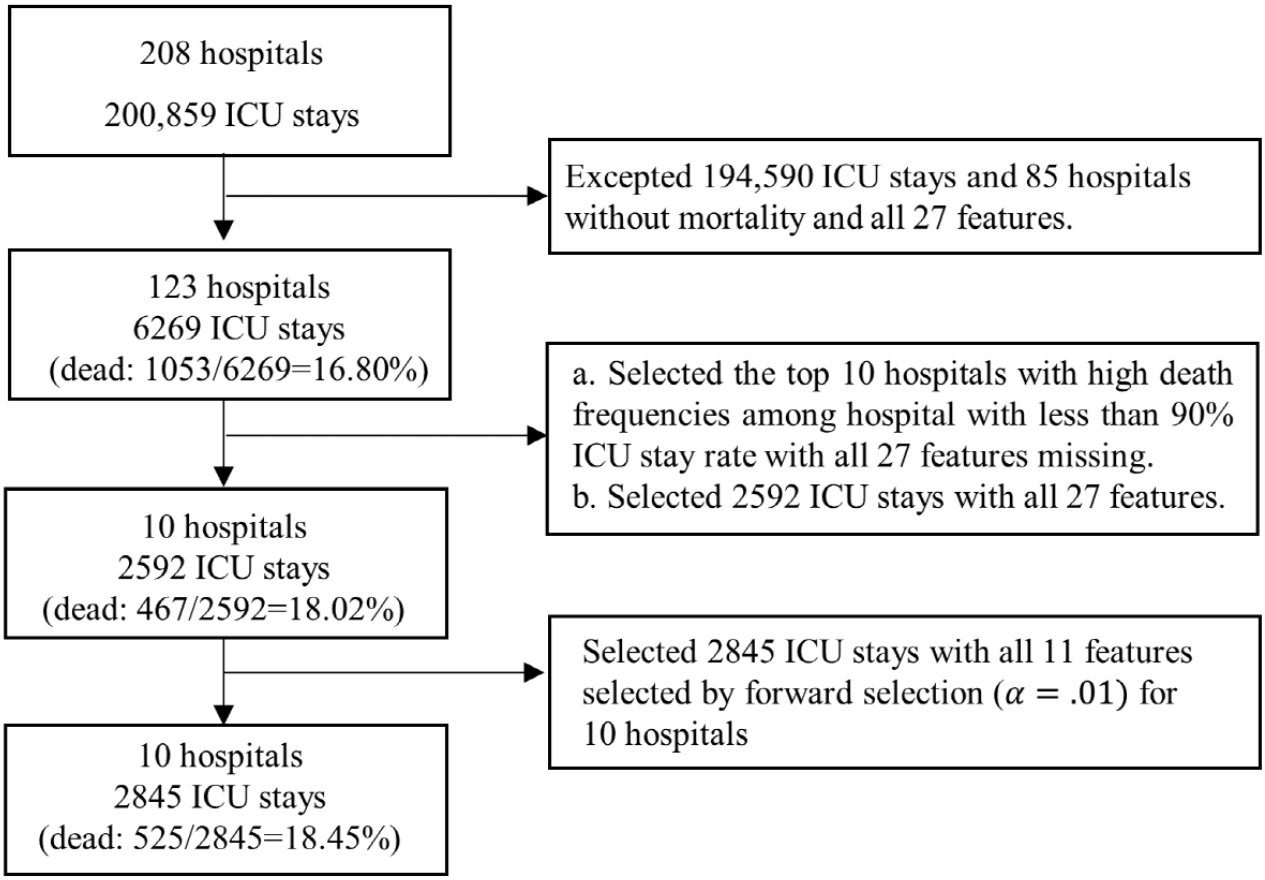
We used the electronic intensive care unit (eICU) Collaborative Research Database [28] to evaluate the validity of the weight model. The eICU Collaborative Research Database is a multi-institution ICU database of eICU programs across the United States, and contains approximately 200,000 admissions to ICUs monitored by 208 hospitals (data collected between 2014 and 2015).

The model to be applied to the weight-based integrated model used a logistic regression model to predict mortality after ICU admission. As features, 27 variables included in the Acute Physiology, Age, and Chronic Health Evaluation (APACHE) classification system were considered. The APACHE score is a severity-of-disease classification system [29], one of several ICU scoring systems. Therefore, we considered 27 variables from the APACHE system as mortality predictors for patients in the ICU. In the eICU database, the APACHE III score was calculated, and the 27 variables used to calculate the score were listed.

We selected 10 hospitals with a total of 2845 ICU stays, out of 208 hospitals with a total of 200,859 ICU stays, as our horizontally partitioned data set (Figure 2). To select the horizontally partitioned data of 10 hospitals, 6269 ICU stays

(123 hospitals) with both mortality and 27 feature values were selected. We selected the top 10 hospitals with higher death frequencies among those having less than 90% ICU stay rate with all 27 features missing. Moreover, 11 features were selected by forward selection (significant level: .01) of 27 features at 2592 ICU stays for 10 hospitals. The selected 11 features were Glasgow Coma Scale score, pH, blood urea nitrogen, fraction of inspired oxygen, temperature, bilirubin, albumin, age, partial pressure of carbon dioxide, partial pressure of oxygen, and pulse rate.

Figure 2. Selection process for hospitals and intensive care unit (ICU) stays.



When developing a predictive model, the number of events compared with the number of predictors is a key factor to determine the performance of the logistic regression model [30]. The models applied to data with low events per variable produce inaccurate and biased results [31]. A total of 10 events per variable are widely used as a criterion for logistic regression models [32,33]. Most hospitals do not satisfy the 10 events per variable criterion based on the 11 features mentioned. Therefore, the firth logistic regression model [34], which can estimate unbiased parameters in data with low event frequencies, was used for accurate parameter sharing between hospitals when applying the weight-based integrated model.

Validation and Evaluation of the Weight-Based Integrated Model

The logistic regression model was used for the simulation data, whereas the firth logistic regression model was used for the real data. To calculate the loss of 2 logistic models, we proceeded according to the process detailed in Figure 1B using the log loss function, –ln L(p). The reciprocal of the log loss risk for all data in each partition model was used as the criterion for calculating the weight. We also used the results of the first simulation as the number of repetitions required to calculate the weight. The ratio of $Z^{(1)}$ to $Z^{(2)}$ was 3:1 for all simulations. In addition, in real data with low event frequency, $Z^{(1)}$ and $Z^{(2)}$ were generated at a 1:1 ratio for both dead and alive cases to build a more stable model in $Z^{(1)}$.

To evaluate the weight-based integrated model, we compared the results of the weight-based integrated model and the centralized model using 10 hospitals from the eICU database, in terms of the ROC curve, AUC, and estimated OR, on the 11 features. In addition, we used the Hosmer–Lemeshow test [35], where $P<.05$ indicates poor calibration, to assess the calibration of the proposed weight-based integrated model and centralized model for central data, along with the 10 models of each hospital.

The comparison of AUCs and ORs between the 2 models was evaluated based on the proportion of overlap of the 95% CIs. The proportion of overlap was defined as the ratio of overlap of two 95% CIs in the margin of error, which is the half-width of the 95% CI of the longer length. If a CI is remarkably short and is included in the other CI to be compared, then the proportion of overlap calculated based on the shorter CI is 2, which is a perfect match between the 2 CIs, regardless of the value of the longer CI. Therefore, the proportion of overlap was

calculated based on the longer CI for a more conservative evaluation criterion. For the independent group *t* test that compares the 2 means, when the proportion of overlap is approximately 0.5 or less, it indicates that the 2-tailed *P* value is less than .05 [36]. We determined that the 2 CIs did not differ significantly at a significance level of .05 when the proportion of overlap was more than 0.5 and confirmed how close the proportion of overlap was to 2.

Based on the results of OR estimation for 11 features, we compared the results of our weight-based integrated model and conventional meta-analysis (for a fixed effect model using the inverse of the variance of the effect estimate as a weight). The meta-analysis is similar to the weight-based integrated model as the OR of a multi-institution is estimated by setting institution-specific weights and averaging the OR of each institution based on the weights, although the method of weight calculation of the meta-analysis varies from the proposed weight-based integrated model. We compared the proportional overlap of 95% CI and the relative bias of point estimates for the centralized model between the weight-based integrated model and the meta-analysis.

To perform external validation, we selected the top 5 hospitals as the external validation hospitals (ie, those with a high mortality rate and less than 90% ICU stay rate with all 27 features missing) after selecting 10 hospitals for the central data. By summarizing the AUC as a result of external validation, we confirmed whether the predictive performance on each external validation hospital in the weight-based integrated model is similar to that of the centralized model. We also evaluated whether the weight-based integrated model ultimately improves the average predictive performance when compared with a model of a single hospital through an average AUC on 5 external validations. In addition, the 3 weighting methods (ie, CS-Avg, n-Avg, and Avg) were applied to external validation and compared with the weight-based integrated model.

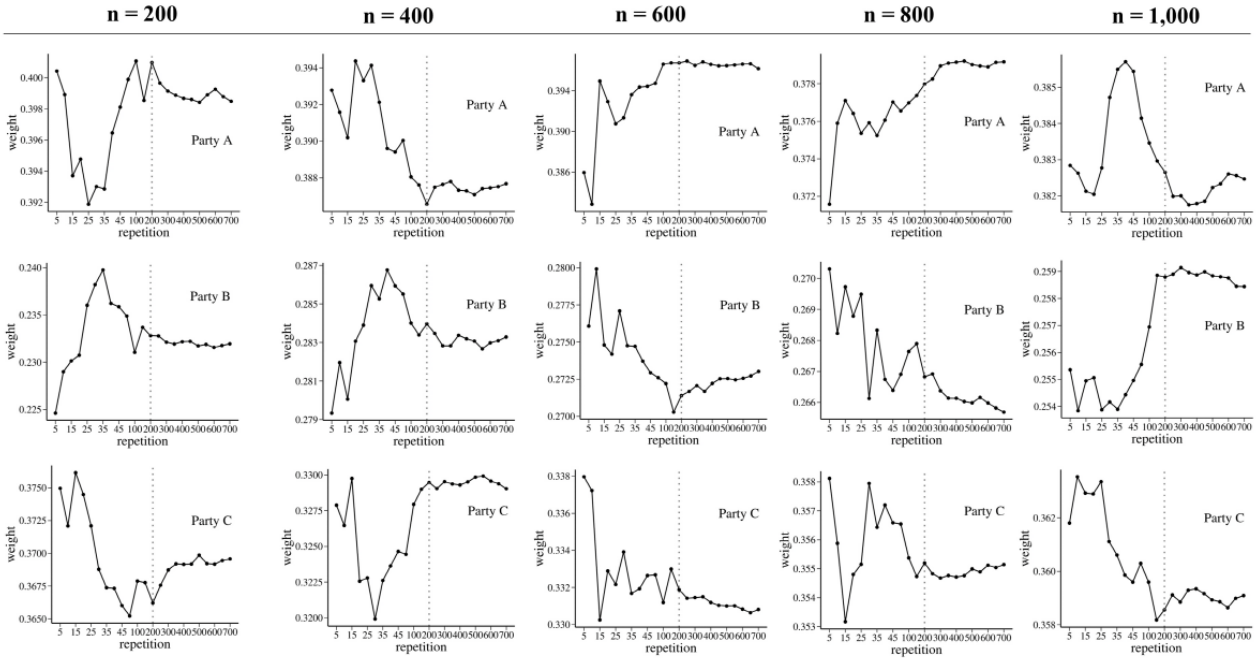
The simulation studies and experiments with real horizontally partitioned data were performed using R 3.6.0 (R Foundation for Statistical Computing).

Results

Simulation 1: Optimal Repetitions m

In simulation 1, to propose optimal repetitions *m* of the weight-based integrated model, the size of each party was simulated as 200, 400, 600, 800, and 1000, and the weight values tended to stabilize as the number of repetitions increased (Figure 3). Moreover, as the data size *n* of each party decreased, the change in the weight pattern according to the number of repetitions became relatively large. For all data size *n*, graphs in Figure 3 showed a relatively flat pattern of weights after 200 repetitions. Therefore, we set *m* to 200. That is, in the second and third simulations, and the experiment using real data, we calculated the weights of each partition model and estimated the parameters of the weight-based integrated model based on 200 repetitions.

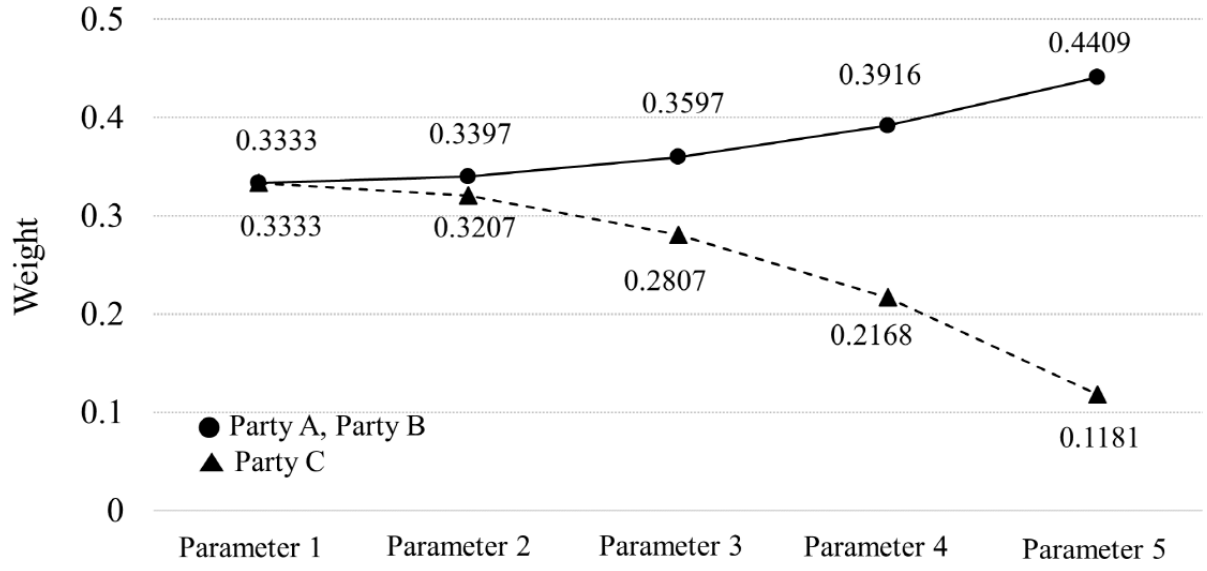
Figure 3. Weights of 3 parties according to the number of repetitions for sizes of 200, 400, 600, 800, and 1000. Vertical lines represent 200 repetitions.



Simulation 2: Features of the Weight Calculated From Weight-Based Integrated Model

To confirm the characteristics of the weights calculated using the weight-based integrated model, party C, among the 3 parties, was considered as a biased party. Figure 4 shows the results of the first scenario to confirm the change of weight according to the goodness of fit. The same weights, 0.3333, are derived for parameter 1 for all parties, where A, B, and C all have the same data. Thereafter, as the degree of bias of party C gradually increases (ie, from parameter 2 to parameter 5), the weight of party C decreases. In other words, under the same data size, the smaller the goodness of fit for the total party of a partition model with different characteristics, the smaller the weight.

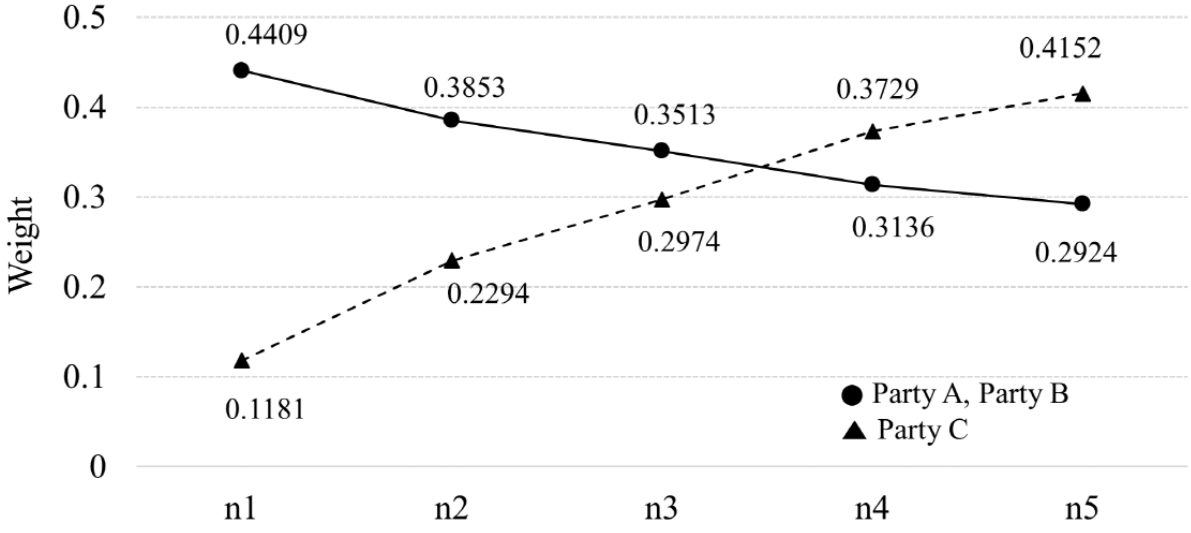
Figure 4. Change pattern of weights according to goodness of fit for central data (scenario 1 of simulation 2), and adjusted parameters for the 5 features of parties A, B, and C with size 1000.



Party A (n=1000)	1	1	1	1	1
Party B (n=1000)	1	1	1	1	1
Party C (n=1000)	1	0.5	−0.5	−1	−2

As shown in the results of Figure 5 (scenario 2), the data size of the biased party C was gradually increased to examine the weight change according to the data size under the setting of parameter 5. When the data size of all 3 parties was equal to 1000, the weight of party C was 0.1181, which was relatively small compared with parties A and B. However, the weight of party C also increased as its data size gradually increased. In particular, after the data size of party C became 4000/6000 (66.67% of the centralized data), the weight of the biased party C became larger than that of the other 2 parties. That is, even in a biased party, the weight can be increased if the ratio of data size to the centralized data increases.

Figure 5. Change pattern of weights according to the ratio of data size to central data (scenario 2 of simulation 2), adjusted data sizes of party C, and ratios of data size to centralized data for parties A, B, and C.



Party A ($all\ \beta$ set 1)	1000 (33.33%)	1000 (25.00%)	1000 (20.00%)	1000 (16.67%)	1000 (14.29%)
Party B ($all\ \beta$ set 1)	1000 (33.33%)	1000 (25.00%)	1000 (20.00%)	1000 (16.67%)	1000 (14.29%)
Party C ($all\ \beta$ set −2)	1000 (33.33%)	2000 (50.00%)	3000 (60.00%)	4000 (66.66%)	5000 (71.42%)

These two results of simulation 2 show that the weights of the weight-based integrated model consider not only the goodness of fit for the central data but also the ratio of data size to the central data.

Simulation 3: Comparative Analysis With Alternative Weighting Methods

Multimedia Appendices 6 and 7 show the comparison results of 200 simulations on the weight of the weight-based integrated model and the other 3 weighting methods (CS-Avg, n-Avg, and Avg). In each simulation setting, we summarized the distribution of 200 average AUC for 4 validation parties, the difference in average AUC between each weighting method and the centralized model, and the average weights of 4 parties (A, B, C, and D), according to 200 simulations.

The results for the first scenario are shown in Multimedia Appendix 6. The data characteristics of parties C and D are gradually heterogeneous with those of party A and B as they go to the left, middle, and right. When the data sizes and characteristics of the 4 parties were all the same (the left in Multimedia Appendix 6), the distributions of the 200 average AUC of each weighting method and the centralized model were almost the same, and the average weights of parties A, B, C, and D were approximately 0.25, which is almost equal. However, as the data characteristics of parties C and D were more different from those of parties A and B (from the left to the right), the predictive performances of the 4 weighting methods were distinctly different. The distribution of the average AUC of CS-Avg showed the largest difference from that of the centralized model, and the weight-based integrated model showed the distribution of average AUC most similar to that of the centralized model. In the first scenario, as the data sizes of the 4 parties were the same, the weights of the 4 parties in both n-Avg and Avg were set equal to 0.25, and the distributions of the average AUC of both weighting methods were the same. As the data characteristics change, the weight-based integrated model and CS-Avg gradually assigned a greater weight to parties C and D. However, as CS-Avg assigned a weight of 0 to one of either A or B, the differences in weight between the 4 parties were greater than that of the weight-based integrated model.

The results for the second scenario are summarized in Multimedia Appendix 7. The data characteristics of the 4 parties were set identically with the condition corresponding to (3) of the first scenario, and the data sizes of parties C and D increased toward the left, middle, and right. Similar to the results of the first scenario, the distribution of the average AUC of the weight-based integrated model was the most similar to that of the centralized model, and the distribution of CS-Avg was the most different. As n-Avg reflects the change in data size, the distribution of average AUC differed from Avg as it goes to the right, and it was closer to the distribution of the centralized model than in the first scenario. As CS-Avg does not reflect the data size, even if the data size of parties C and D increased, the weights of the 4 parties remained almost unchanged. However, the weight-based integrated model gradually provided large weights to parties C and D with large data sizes. Furthermore, as n-Avg reflects the data size, but does not reflect the data characteristics, there was a difference from the weight of the weight-based integrated model reflecting both. Avg assigned 4 parties a fixed weight of 0.25 under any conditions.

Validation Results on Horizontally Partitioned eICU Data

A total of 2845 ICU stays (dead: 525, alive: 2320) were arranged from 10 hospitals. Among the 2845 ICU stays, the total of $Z^{(1)}$ of the entire hospital was 1430 ICU stays, and the total of $Z^{(2)}$ was 1415 ICU stays (refer to Multimedia Appendix 1). Table 1 presents the results of AUC from the firth logistic regression model in each of the 10 hospitals. The predictive power of the models from each hospital differs from the smallest predictive power of 80.93% (hospital 6) to the largest predictive power of 92.00% (hospital 10).

The 200 log loss values for the total $Z^{(2)}$ (n=1415) of each hospital model and the final weights of each hospital model were calculated from 200 repetitions (Table 1). A large distribution of loss in a hospital indicates that the goodness of fit of the hospital model is not good for all data from 10 hospitals. Therefore, the weight of a hospital with a relatively small loss distribution was calculated to be small. Further, a hospital with a small ratio of data size to central data (2845 ICU stays) tends to have a small weight. For example, in hospital 1, the distribution of the loss is the smallest, and the ratio of data size to the central data is the largest (510/2845, 17.93%). Therefore, the largest weight of 0.1188 was assigned to hospital 1. Conversely, hospital 10 has the largest distribution of loss, and the ratio of data size to the central data is the smallest (125/2845, 4.39%). Therefore, the smallest weight of 0.0583 was assigned to hospital 10. Hospitals 3 and 4 were given the same weight of 0.1109. However, the ratio of data size to central data in hospital 3 (268/2845, 9.42%) was smaller than that of hospital 4 (338/2845, 11.88%), and the loss distribution tended to be slightly smaller for hospital 3. As observed in the results of simulation 2, the weight of the weight-based integrated model is affected by both the ratio of the central data and the goodness of fit to the central data.

Table 1. AUC, log loss, and weights for 10 models of each institution (N=2845).

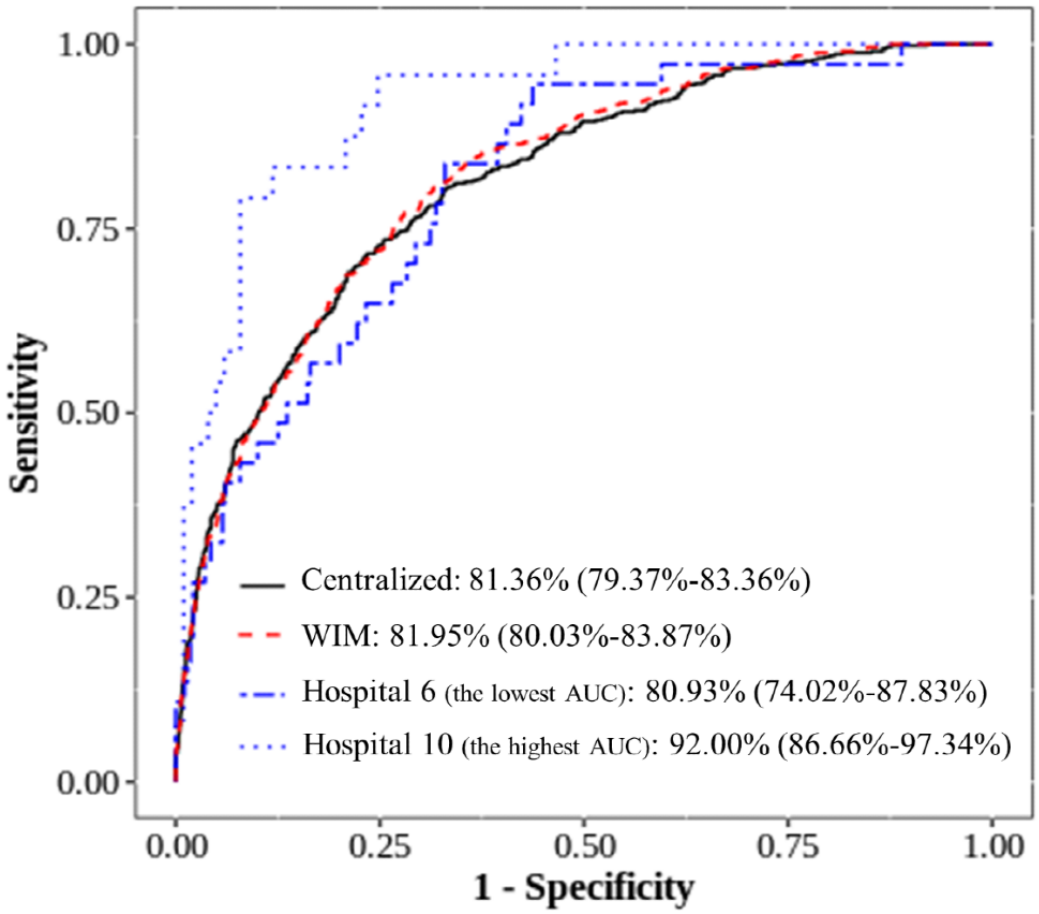
Hospital number	n/N (%)	AUC[a] (95% CI)	Log loss from 200 repetitions		Weight
			Median	(Min, Max)	
1	510/2845 (17.93)	83.81% (79.99%-87.63%)	575.18	(535.45, 668.13)	0.1188
2	387/2845 (13.60)	82.14% (76.82%-87.47%)	577.40	(536.59, 754.68)	0.1181
3	268/2845 (9.42)	86.67% (81.57%-91.78%)	616.63	(547.65, 755.15)	0.1109
4	338/2845 (11.88)	86.48% (81.43%-91.53%)	617.14	(552.61, 787.62)	0.1109
5	231/2845 (8.12)	86.29% (80.19%-92.4%)	723.90	(572.31, 1814)	0.0929
6	316/2845 (11.11)	80.93% (74.02%-87.83%)	626.65	(539.71, 978.16)	0.1076
7	308/2845 (10.83)	85.95% (78.23%-93.67%)	665.89	(561.92, 1071.16)	0.1024
8	197/2845 (6.92)	83.81% (75.88%-91.73%)	712.29	(569.31, 7280.35)	0.0912
9	165/2845 (5.79)	86.63% (79.2%-94.05%)	758.66	(566.39, 1774.99)	0.0890
10	125/2845 (4.39)	92% (86.66%-97.34%)	1008.64	(634.35, 13,722.49)	0.0583

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

The Hosmer–Lemeshow goodness-of-fit test demonstrated that the weight-based integrated model and the centralized model fit the central data well, and the 10 models of each hospital fit the data of each hospital well (all P>.05; Multimedia Appendix 3).

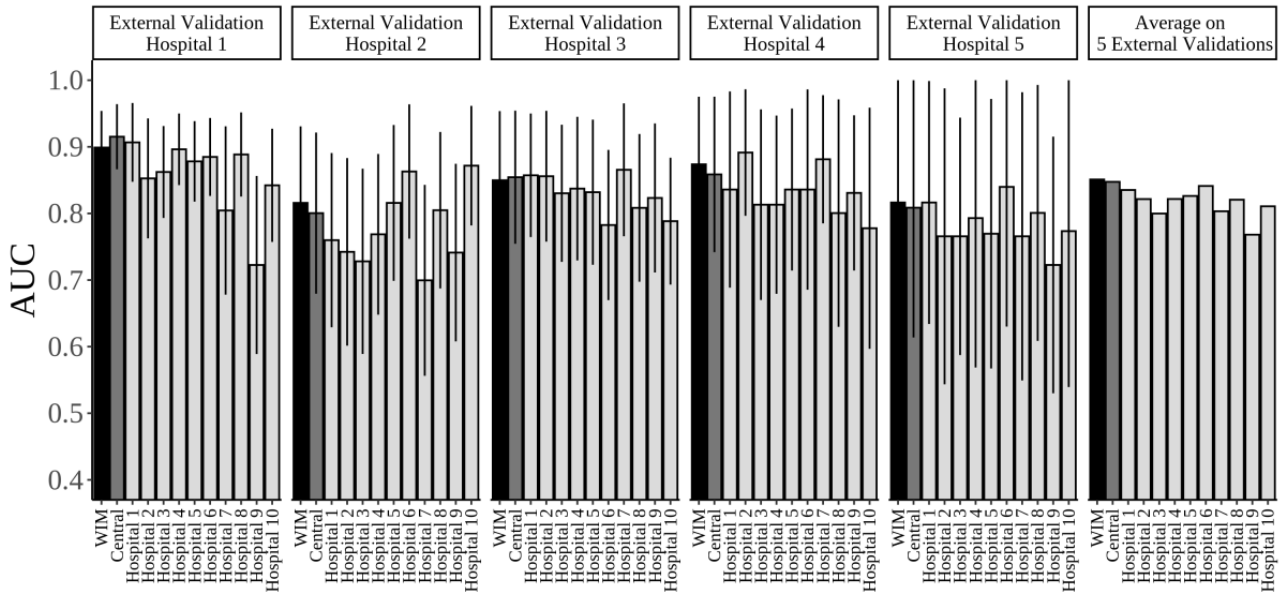
Figure 6 shows the ROC and AUC of the 2 models, the weight-based integrated model and the centralized model based on the central data (2845 stays), and of the 2 hospitals, hospital 6 with the lowest AUC and hospital 10 with the highest AUC (based on the data of each hospital). It was confirmed that the patterns of ROC curves for both the weight-based integrated model and the centralized model are almost the same. The estimated AUC values and 95% CIs were 81.36% (79.37%-83.36%) and 81.95% (80.03%-83.87%) in the centralized model and the weight-based integrated model, respectively (Figure 6). The proportion of overlap of CIs for AUC in both the weight-based integrated model and the centralized model was approximately 1.70. This value is much larger than 0.5, which is the level that we consider to indicate a significant difference at a significance level of .05, and is close to 2, which is the criterion for completely matching 2 CIs. Therefore, the calculated CIs for the AUC in both models were almost equal. The model of hospital 10 with the largest AUC was an overfitted model with an AUC 10% greater than for the 2 models (the weight-based integrated model and the centralized model) and the model of hospital 6 did not show much difference in the AUC value compared with the 2 models.

Figure 6. Area under the receiver operating characteristic curve (AUC), log loss from 200 repetitions, and weights. WIM: weight-based integrated model.



A total of 535 ICU stays were selected as the 5 external validation hospitals. The frequency and rate of mortality of external validation hospitals 1, 2, 3, 4, and 5 were 20/155 (12.9%), 19/67 (28.36%), 24/226 (10.62%), 11/47 (23.4%), and 8/40 (20%), respectively. Figure 7 shows the AUC of each external validation hospital and the average AUC on 5 external validations. Multimedia Appendix 4 presents the values of the AUC (95% CI) shown in Figure 7, as well as the proportional overlap for the 95% CI of the weight-based integrated model and the centralized model. The weight-based integrated model had similar predictive performances to the centralized model in 5 external validations. In each external validation, the proportional overlap of the 95% CI for the centralized model and the weight-based integrated model was 1.59, 1.82, 1.92, 1.74, and 1.93 for external validation hospitals 1, 2, 3, 4, and 5, respectively. In addition, the average AUC was 84.74% and 85.09% for the centralized model and the weight-based integrated model, respectively. In each of the 5 external validation hospitals, a model of a single hospital out of 10 models showed higher AUC than the weight-based integrated model. However, the weight-based integrated model demonstrated the highest average predictive performance on the 5 external validation hospitals (Figure 7).

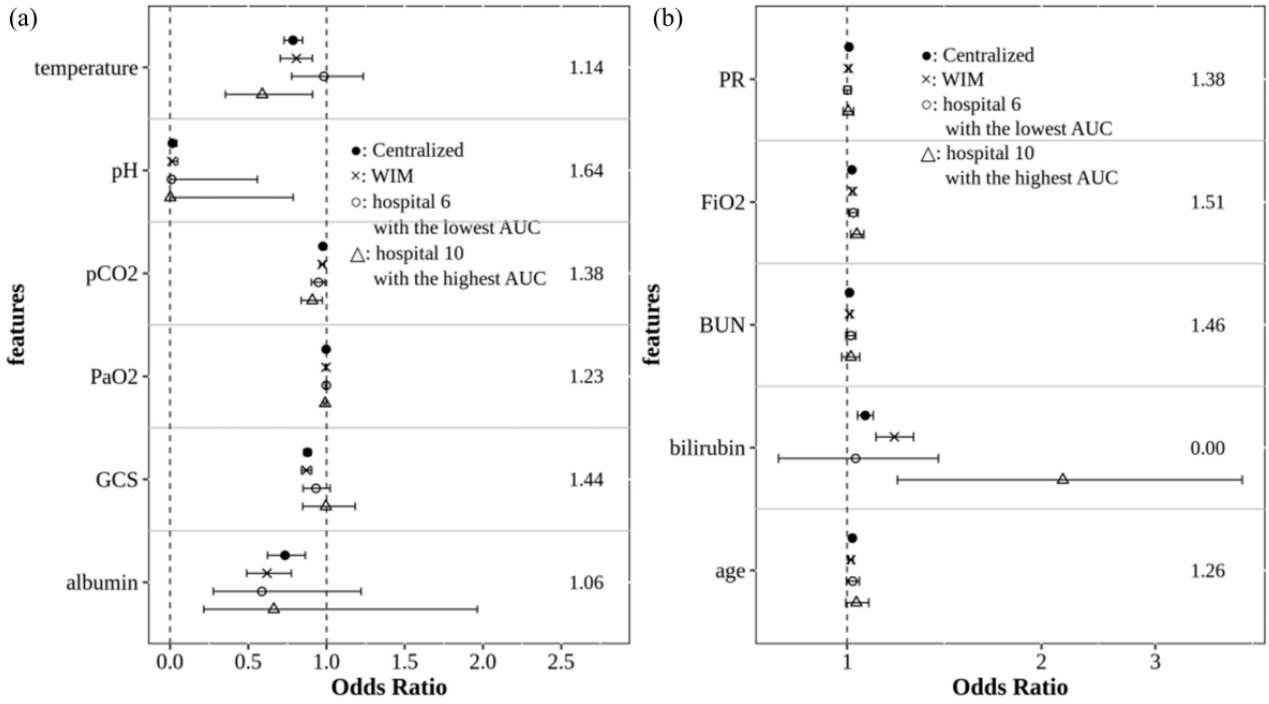
Figure 7. Results of AUC of external validation for the centralized model, the WIM, and 10 models of each hospital (error bar: 95% CI). Black, dark gray, and light gray indicate WIM, centralized model, and 10 models of each hospital, respectively. AUC: area under the receiver operating characteristic curve; WIM: weight-based integrated model.



Multimedia Appendix 8 shows the comparison results of the external validation of the weight-based integrated model and 3 other weighting methods, namely, CS-Avg, n-Avg, and Avg. The proportional overlaps of the 95% CI on the AUC of the 3 weighting methods were also high, similar to those of the weight-based integrated model. In addition, the average AUCs on the 5 external validation hospitals for each weighting method were similar to each other (weight-based integrated model, 0.8509; CS-Avg, 0.8519; n-Avg, 0.8502; Avg, 0.8507).

Figure 8 shows the OR and 95% CI of 11 features estimated using the weight-based integrated model and the centralized model, based on the central data (2845 stays), and 2 hospitals (hospital 6 with the lowest AUC and hospital 10 with the highest AUC). The 11 features were significant in both the centralized model and the weight-based integrated model, and the direction of OR significance was consistent in both models. Figure 8A presents the result of significant features with OR < 1, whereas Figure 8B presents the result of significant features with OR > 1. For the proportional overlap of 95% CI of OR between the weight-based integrated model and the centralized model, all 10 features, except bilirubin, showed a result exceeding 1 (significant difference is 0.5 at a significance level of .05), and the ORs estimated in the 2 models did not differ significantly. In bilirubin, 95% CI of the 2 models did not overlap. For each of the 11 features, ORs were estimated differently in the 10 hospitals, including hospitals 6 and 10 indicated in the graph (refer to Multimedia Appendix 2). The ORs estimated using the weight-based integrated model showed most similar estimation results to the centralized model, compared with the ORs estimated from each hospital model.

Figure 8. Comparison of estimated OR and 95% CI on 11 features in the firth logistic regression model: (A) features with OR < 1 and (B) features with OR > 1. The numbers on the right sides of the figures are the proportional overlap of 95% CI of OR between the WIM and the centralized model. AUC: area under the receiver operating characteristic curve; BUN: blood urea nitrogen; FiO2: fraction of inspired oxygen; GCS: Glasgow Coma Scale; OR: odds ratio; PaO2: partial pressure of oxygen; pCO2: partial pressure of carbon dioxide; PR: pulse rate; WIM: weight-based integrated model.



As a result of the comparison with the meta-analysis, depending on the feature, the degree of similarity to the centralized model was slightly different between the weight-based integrated model and the meta-analysis in terms of the proportional overlap of 95% CI and relative bias (Multimedia Appendix 5). Based on the criteria of the proportional overlap of 95% CI, the overlap of the weight-based integrated model and the meta-analysis for pH was 1.64 and 1.33, respectively. For Glasgow Coma Scale, pH, temperature, and partial pressure of carbon dioxide, the relative bias of the weight-based integrated model was smaller than that of the meta-analysis. These results indicate that the weight-based integrated model was closer to the centralized model than the meta-analysis. However, bilirubin, whose proportional overlap was 0 in the weight-based integrated model, showed a proportional overlap of 1.69 in the meta-analysis. In addition, the relative bias of bilirubin was 10.94% and 0.66% in the weight-based integrated model and the meta-analysis, respectively.

Discussion

Principal Findings

The proposed model (the weight-based integrated model) was developed to build an integrated predictive model from horizontally partitioned data without requiring physical data sharing. The weight-based integrated model is an algorithm that does not require an iterative process and can extend the model to be applied by introducing the concept of a flexible weight of a partition model. Unlike previous methodologies of building a model of central data under privacy-preserving conditions, the proposed model has the following novelties.

First, the weight-based integrated model does not require iterative communication to construct a model that approximates the centralized model. The methods that use distributed computing require an iterative exchange of information between the institutions and the central server, which is time consuming and labor intensive in practice [20]. This practical limitation can be a barrier to the application of distributed algorithms in a research consortium [20]. In cross-silo federated learning [8] with an iterative process, all clients are always available and should participate in each iteration. In other words, if a party is not available in the middle of the iteration process, the entire process is stopped. Conversely, the weight-based integrated model can build an integrated model by adjusting the weights even if a party becomes unavailable during the process. In terms of communication efficiency, naïve application of previous methodologies can yield procedures that incur exorbitant communication costs [37].

Second, the weight of the weight-based integrated model is a flexible weight derived from 2 factors, data size and the goodness of fit of each party's model to the entire data (Figures 4 and 5). As the ratio of the data sizes of each party in the central data increases, the partition model would be closer to the centralized model. Therefore, the data size should be considered in the weighting of the partition model. If the partition model fits well to the central data, then it would be a model that describes the central data well. Therefore, the goodness of fit should also be considered with the data size. A key characteristic of the weight-based integrated model is that the weight of each partition model is derived by considering these 2 factors simultaneously. In addition, when constructing the weight-based integrated predictive model in the weight-based integrated model, the weights of the model of each party are generated *m*

times (Figure 1), and the average of *m* weights is set as the final weight of the model of the party. Therefore, depending on how *m* is set, the final weights of the models of each party vary. In simulation 1, we found the optimal *m*, where the final weight remained almost unchanged while increasing the size of *m* under various data sizes of the 3 parties. The results showed that there was little change in the final weight when *m* exceeded 200 for all data sizes of the parties (Figure 3).

Third, the weight-based integrated model is a flexible algorithm in terms of scalability of the model to be applied. As the proposed model builds each partition model independently and then integrates them based on the weight, it only needs to change the form of parameters in step 2 and the loss function in step 3, depending on the model.

Validation and Evaluation of the Weight-Based Integrated Model

We evaluated the validity of the weight-based integrated model in terms of predictive power and parameter estimation, compared with the centralized model. Experimental results using real horizontally partitioned data demonstrated that the weight-based integrated model provides a close approximation to the centralized model and improves the average predictive performance.

In terms of predictive power, the weight-based integrated model was substantially similar to the centralized model based on the results of the ROC curve and AUC. The weight-based integrated model provided a weighted average model by integrating each partition model overfitted or underfitted, compared with the centralized model (Figure 6). The multi-institutional predictive model aims to develop a generalized model that can improve the predictive performance for the data that were not used in the model. To confirm whether the proposed model satisfies this objective, we selected 5 hospitals that were not used in the weight-based integrated model and performed an external validation. Consequently, for the estimation of the AUC for each external validation hospital, the weight-based integrated model exhibited almost similar results as the centralized model. In addition, its average AUC for the 5 external validation hospitals was higher than that of the 10 models of each hospital (Figure 7, Multimedia Appendix 4).

In terms of parameter estimation, based on the results of the proportional overlap (0.5 or less indicates a significant difference at a significance level of .05; 2 indicates two CIs overlapping completely) for 95% CI of OR (Figure 8), 10 features were over 1 or 1.5. The results of parameter estimation between the weight-based integrated model and the centralized model were quite similar. However, the 95% CI of bilirubin did not overlap between the 2 models; the estimation of bilirubin was different at the significance level of 5%. As observed in the 95% CI of 10 models on each hospital for bilirubin (refer to Multimedia Appendix 2), hospital 5 with a weight of 0.0929 and hospital 10 with a weight of 0.0583 had no overlap with the centralized model. The reason that the OR for bilirubin of the weight-based integrated model differed from the centralized model is that the proportional overlap of hospital 5 with large weight was 0. Further, the estimated OR from hospital 10 was unstable and biased compared with other hospitals. The OR and 95% CI for bilirubin of the centralized model and the weight-based integrated model were 1.07 (1.04-1.10) and 1.18 (1.11-1.27), respectively (Multimedia Appendix 2). Although the 95% CI of the weight-based integrated model did not overlap with the centralized model, in the 2 models, the statistical significance of OR and the direction of interpretation are consistent, and the overall CI of the weight-based integrated model is not far off from that of the centralized model, compared with CIs of 10 hospital models.

The results of comparison with the meta-analysis in experiments using real data indicate that, for the OR estimates of 4 out of 11 features, the relative biases of the weight-based integrated model were slightly less than those of the meta-analysis. The weight-based integrated model generally showed similar results to the meta-analysis in terms of estimation of ORs. However, depending on the features, owing to the difference in weight calculation between the meta-analysis and the weight-based integrated model, there were differences in proportional overlap of 95% CI and relative bias. The weight of the meta-analysis has institution-specific characteristics. However, as it is adjusted based on the variance of an estimator of OR, the different weights are generated even for the same institution depending on which feature's OR is estimated. By contrast, as the weights in our proposed weight-based integrated model are assigned to the model of each institution, even if the features to be estimated are different, the same weight is given to the same institution. Although the weight of the meta-analysis has feature-specific characteristics more than the weight of the weight-based integrated model, it does not represent the weight for a model of an institution unlike the weight-based integrated model. Therefore, it cannot be regarded as a weight that encompasses the purpose of building a predictive model.

When applying the weight-based integrated model, it is necessary to consider the following: To calculate the weight of each institution in the weight-based integrated model, the data of each institution is divided into $Z^{(1)}$, for building the model of each institution, and $Z^{(2)}$,for measuring the predictive performances of the models of all institutions. If the data size (especially the frequency of outcome of interest) of an institution is insufficient, the model of the institution generated by $Z^{(1)}$ will be unstable, and it will be difficult to accurately calculate the predictive performance from $Z^{(2)}$. Therefore, the data size of each institution should be sufficient to divide them into $Z^{(1)}$ and $Z^{(2)}$. In addition, based on the results of the external validation, the predictive performances of each of the 5 external validation hospitals were better in the model of single hospitals, compared with those of the weight-based integrated model. In other words, the weight-based integrated model may not be a good option for the purpose of improving the predictive performance of a specific hospital (of the 5 hospitals). By contrast, as the purpose for improving the average predictive performance of the 5 hospitals, the weight-based integrated model can provide a robust unified model. In our experiment using real data, the weight-based integrated model showed the best average predictive performance on 5 external validation hospitals. However, there may be cases where the weight-based integrated model does not show the best average predictive performance.

For example, when a relatively heterogeneous model among the hospitals included in the weight-based integrated model exists, and the hospital exhibits heterogeneous characteristics toward all external hospitals, if the predictive performance of the model of the heterogeneous hospital in all external validation hospitals is low, the average predictive performance of the weight-based integrated model may be poor. As the weight-based integrated model averages the models of each hospital based on the weight, the overall prediction performance may be low owing to the inclusion of a heterogeneous hospital with poor predictive performance for external validation hospitals, although it is given a small weight in the weight-based integrated model. To avoid this case, it is necessary to form hospitals of the weight-based integrated model to ensure that the overall characteristics of the hospitals in which the weight-based integrated model will be applied are evenly reflected.

The weight-based integrated model is a similar algorithm to the MCCG [23,24], as it does not require an iterative communication process between institutions and constructs a generalized predictive model by integrating the models of each institution based on the weights per institution. However, the generalization process of both models varies. The weight of the weight-based integrated model is calculated by measuring the heterogeneity of the predictive performance for the central data of the models per institution in order to estimate the centralized model. Conversely, the weight of the MCCG is calculated by measuring the heterogeneity of the predictive performance for a specific target institution of the models of the source institutions used to develop the multi-institutional predictive model in order to improve the predictive performance of the target institution. Owing to this difference in the weight calculation method, the weight-based integrated model provides a generalized model by building a unified model that reflects all the characteristics of multiple institutions, whereas the MCCG provides a generalized model by changing the model through weight adjustments according to the target hospital. In the weight-based integrated model, communication occurs between institutions only once during the process of the algorithm. Conversely, the MCCG requires communication whenever the target institution changes as communication occurs between the source and target institutions. In particular, if the goal is to build a single unified predictive model to be applied to multiple institutions, the weight-based integrated model can provide a robust model. However, if the goal is to build a predictive model for a specific target institution, the MCCG can provide a better model. Therefore, an algorithm should be strategically selected according to the goal.

Comparison With Other Weighting Methods

We demonstrated the characteristics of the weight of the weight-based integrated model through comparative analysis with other comparable weighting methods (CS-Avg, n-Avg, and Avg) [26]. The weight of the weight-based integrated model has characteristics that are calculated by considering the data size of each party and the predictive performance of central data consisting of all parties, and these characteristics were clearly distinguished from other weights, as shown in the third simulation study (Multimedia Appendix 6).

In the weight-based integrated model, the weights were adjusted as the data characteristics of the parties changed under the same data size, and the weights were adjusted as the data sizes of the parties changed under the same data characteristics. By contrast, Avg always assigned a fixed weight that does not reflect the different characteristics and data sizes of each party, and n-Avg assigned a weight that reflects only the change in the data size of each party. In addition, CS-Avg did not reflect the change in data size, but rather reflected the change in data characteristics between parties. Because CS-Avg assigns a weight of 0 to a party with the lowest performance to other parties, the party with a weight of 0 was not considered in the model. Therefore, compared with other weights, the predictive performance of CS-Avg was the most different from that of the centralized model. The weight of the weight-based integrated model distinguished from other weights reflects the characteristics of each party in the central data in terms of data size and data characteristics of each party. The weight-based integrated model with these characteristics can build a model that shows similar predictive performance as the centralized model, compared with other weighting methods.

In our experiment using real data, there were few differences in the results of external validation between the weight-based integrated model and other weighting methods as the weights assigned to the 10 hospitals differed only slightly for each weighting method (Multimedia Appendix 8). The characteristics of each weighting method were not revealed in the application of real data. However, it can be confirmed that, through a third simulation study, a difference exists in the concept from which the weight of each weighting method is derived, and the weight of the weight-based integrated model has a characteristic for estimating the centralized model.

Limitations

It was mentioned that the weight-based integrated model is a model without an iterative process as the novelty. However, we did not evaluate its efficiency due to the absence of iterative processes in the real distributed environment. In addition, this study verified the proposed method using 2 logistic regression models, and we did not confirm the validity of the weight-based integrated model by applying other models. As shown in the results of the estimated OR for bilirubin in Figure 7, when estimating the parameters in the weight-based integrated model, inaccurate information can be provided, compared with the centralized model. As the parameters of the proposed method were estimated by assigning weights to each party's coefficient, the parameter estimation can be influenced by the characteristics of a specific party. This limitation indicates that when a feature is estimated to be highly biased in one party, and the weight of the party is not small relative to another, it needs to interpret the estimated value carefully from the weight-based integrated model. In the future, we will explore the application and efficiency of the weight-based integrated model in a real distributed environment based on a model that has not been applied in this study.

Conclusions

In this study, we developed a weight-based integrated model, which can build an integrated predictive model with noniterative

communication between institutions. The weight-based integrated model, which uses the concept of weights for each institution, is a privacy-protecting analytic method that can reduce the burden of distributed computing and improve the average predictive performance of external validation institutions. The proposed weight-based integrated model can provide an efficient distributed research algorithm to improve the usage of multi-institutional data.

Acknowledgments

This research was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI19C1015).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The frequency and rate of events for each of total, Z(1) and Z(2) in 10 hospitals.
[DOCX File , 21 KB - medinform_v9i4e21043_app1.docx]

Multimedia Appendix 2
Estimated OR in the centralized model, the weight-based integrated model, and 10 models of each hospital in experiments using real data.
[PDF File (Adobe PDF File), 445 KB - medinform_v9i4e21043_app2.pdf]

Multimedia Appendix 3
Hosmer-Lemeshow goodness-of-fit tests to assess the calibration of the weight-based integrated model and centralized model for central data, and the 10 models of each hospital.
[DOCX File , 17 KB - medinform_v9i4e21043_app3.docx]

Multimedia Appendix 4
Average AUC for 5 external validation hospitals and AUC (95% CI) of each external validation hospital in the centralized model, the weight-based integrated model, and 10 models of each of the 10 hospitals.
[DOCX File , 21 KB - medinform_v9i4e21043_app4.docx]

Multimedia Appendix 5
Comparison results of the OR (95% CI) of 11 features between the weight-based integrated model and the meta-analysis.
[DOCX File , 20 KB - medinform_v9i4e21043_app5.docx]

Multimedia Appendix 6
Results of the simulation study for comparison with other weighting methods according to the change of data characteristics under the same data size.
[DOCX File , 580 KB - medinform_v9i4e21043_app6.docx]

Multimedia Appendix 7
Results of the simulation study for comparison with other weighting methods according to the change of data size under the same data characteristics.
[DOCX File , 621 KB - medinform_v9i4e21043_app7.docx]

Multimedia Appendix 8
Results of comparative analysis of external validation by the weighting methods using the eICU data.
[DOCX File , 21 KB - medinform_v9i4e21043_app8.docx]

References

1. Kukull WA, Ganguli M. Generalizability: The trees, the forest, and the low-hanging fruit. Neurology 2012 Jun 04;78(23):1886-1891. [doi: 10.1212/wnl.0b013e318258f812]
2. Katzan IL, Rudick RA. Time to integrate clinical and research informatics. Sci Transl Med 2012 Nov 28;4(162):162fs41-162fs41. [doi: 10.1126/scitranslmed.3004583] [Medline: 23197569]

3. Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, pSCANNER team. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. J Am Med Inform Assoc 2014;21(4):621-626 [FREE Full text] [doi: 10.1136/amiajnl-2014-002751] [Medline: 24780722]
4. Schilling LM, Kwan BM, Drolshagen CT, Hosokawa PW, Brandt E, Pace WD, et al. Scalable Architecture for Federated Translational Inquiries Network (SAFTINet) Technology Infrastructure for a Distributed Data Network. EGEMS (Wash DC) 2013 Jul 10;1(1):1027 [FREE Full text] [doi: 10.13063/2327-9214.1027] [Medline: 25848567]
5. McCarty CA, Chisholm RL, Chute CG, Kullo IJ, Jarvik GP, Larson EB, eMERGE Team. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med Genomics 2011 Jan 26;4(1):13 [FREE Full text] [doi: 10.1186/1755-8794-4-13] [Medline: 21269473]
6. Shi H, Jiang C, Dai W, Jiang X, Tang Y, Ohno-Machado L, et al. Secure Multi-pArty Computation Grid LOgistic REgression (SMAC-GLORE). BMC Med Inform Decis Mak 2016 Jul 25;16(S3):89 [FREE Full text] [doi: 10.1186/s12911-016-0316-1] [Medline: 27454168]
7. Prasser F, Kohlmayer F, Kuhn KA. Efficient and effective pruning strategies for health data de-identification. BMC Med Inform Decis Mak 2016 Apr 30;16(1):49 [FREE Full text] [doi: 10.1186/s12911-016-0287-2] [Medline: 27130179]
8. Feng Y, Yang X, Fang W. Practical and bilateral privacy-preserving federated learning. arXiv 2021.
9. Kairouz EBP, McMahan HB. Advances and Open Problems in Federated Learning. FNT in Machine Learning 2021;14(1):02406503. [doi: 10.1561/2200000083]
10. Adby P. Introduction to Optimization Methods. Berlin, Germany: Springer Science & Business Media; 2013.
11. Wu Y, Jiang X, Kim J, Ohno-Machado L. Grid Binary LOgistic REgression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 2012 Sep 01;19(5):758-764 [FREE Full text] [doi: 10.1136/amiajnl-2012-000862] [Medline: 22511014]
12. Wu Y, Jiang X, Wang S, Jiang W, Li P, Ohno-Machado L. Grid multi-category response logistic models. BMC Med Inform Decis Mak 2015 Mar 18;15(1):10 [FREE Full text] [doi: 10.1186/s12911-015-0133-y] [Medline: 25886151]
13. Lu C, Wang S, Ji Z, Wu Y, Xiong L, Jiang X, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015 Nov;22(6):1212-1219 [FREE Full text] [doi: 10.1093/jamia/ocv083] [Medline: 26159465]
14. Kalbfleisch J, Prentice R. The Statistical Analysis of Failure Time Data. Hoboken, NJ: John Wiley & Sons; 2011.
15. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials 1986 Sep;7(3):177-188. [doi: 10.1016/0197-2456(86)90046-2]
16. Boland M, Parhi P, Li L, Miotto R, Carroll R, Iqbal U, et al. Uncovering exposures responsible for birth season - disease effects: a global study. J Am Med Inform Assoc 2018 Mar 01;25(3):275-288 [FREE Full text] [doi: 10.1093/jamia/ocx105] [Medline: 29036387]
17. Duke JD, Ryan PB, Suchard MA, Hripcsak G, Jin P, Reich C, et al. Risk of angioedema associated with levetiracetam compared with phenytoin: Findings of the observational health data sciences and informatics research network. Epilepsia 2017 Aug 06;58(8):e101-e106 [FREE Full text] [doi: 10.1111/epi.13828] [Medline: 28681416]
18. Vashisht R, Jung K, Schuler A, Banda JM, Park RW, Jin S, et al. Association of Hemoglobin A1c Levels With Use of Sulfonylureas, Dipeptidyl Peptidase 4 Inhibitors, and Thiazolidinediones in Patients With Type 2 Diabetes Treated With Metformin: Analysis From the Observational Health Data Sciences and Informatics Initiative. JAMA Netw Open 2018 Aug 03;1(4):e181755 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.1755] [Medline: 30646124]
19. Ryan PB, Buse JB, Schuemie MJ, DeFalco F, Yuan Z, Stang PE, et al. Comparative effectiveness of canagliflozin, SGLT2 inhibitors and non-SGLT2 inhibitors on the risk of hospitalization for heart failure and amputation in patients with type 2 diabetes mellitus: A real-world meta-analysis of 4 observational databases (OBSERVE-4D). Diabetes Obes Metab 2018 Nov 25;20(11):2585-2597 [FREE Full text] [doi: 10.1111/dom.13424] [Medline: 29938883]
20. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012 Jan 01;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
21. Duan R, Boland MR, Moore JH, Chen Y. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. Pacific Symposium on Biocomputing 2019. 2019. URL: https://psb.stanford.edu/psb-online/proceedings/psb19/duan.pdf [accessed 2021-03-28]
22. Duan R. Learning from local to global-an efficient distributed algorithm for modeling time-to-event data. bioRxiv 2021:-1036. [doi: 10.1101/2020.03.04.977298]
23. Tian Y, Shang Y, Tong D, Chi S, Li J, Kong X, et al. POPCORN: A web service for individual PrognOsis prediction based on multi-center clinical data CollabORatioN without patient-level data sharing. J Biomed Inform 2018 Oct;86:1-14 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.008] [Medline: 30103028]
24. Tian Y, Chen W, Zhou T, Li J, Ding K, Li J. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: A pilot study. Int J Med Inform 2020 Sep;141:104173. [doi: 10.1016/j.ijmedinf.2020.104173] [Medline: 32531725]
25. Brownlee J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. San Juan, PR: Machine Learning Mastery Pty. Ltd; 2020.

26. Patil P, Parmigiani G. Training replicable predictors in multiple studies. Proc Natl Acad Sci U S A 2018 Mar 13;115(11):2578-2583 [FREE Full text] [doi: 10.1073/pnas.1708283115] [Medline: 29531060]
28. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data 2018 Sep 11;5(1):180178 [FREE Full text] [doi: 10.1038/sdata.2018.178] [Medline: 30204154]
29. Le Gall J. A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study. JAMA 1993 Dec 22;270(24):2957. [doi: 10.1001/jama.1993.03510240069035]
30. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Regression modelling strategies for improved prognostic prediction. Stat Med 1984 Apr;3(2):143-152. [doi: 10.1002/sim.4780030207] [Medline: 6463451]
31. Jewell NP. Small-Sample Bias of Point Estimators of the Odds Ratio from Matched Sets. Biometrics 1984 Jun;40(2):421. [doi: 10.2307/2531395]
32. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct 14;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
33. van Smeden M, de Groot JAH, Moons KGM, Collins GS, Altman DG, Eijkemans MJC, et al. No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. BMC Med Res Methodol 2016 Nov 24;16(1):163 [FREE Full text] [doi: 10.1186/s12874-016-0267-3] [Medline: 27881078]
34. Doerken S, Avalos M, Lagarde E, Schumacher M. Penalized logistic regression with low prevalence exposures beyond high dimensional settings. PLoS One 2019 May 20;14(5):e0217057 [FREE Full text] [doi: 10.1371/journal.pone.0217057] [Medline: 31107924]
35. Lemeshow S, Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. Am J Epidemiol 1982 Jan;115(1):92-106. [doi: 10.1093/oxfordjournals.aje.a113284] [Medline: 7055134]
36. Geoff C, Fiona F. Interval estimates for statistical communication: Problems and possible solutions. IASE Satellite. 2005. URL: https://iase-web.org/documents/papers/sat2005/cumming.pdf?1402524993 [accessed 2021-03-28]
37. Jordan MI, Lee JD, Yang Y. Communication-Efficient Distributed Statistical Inference. Journal of the American Statistical Association 2018 Nov 13;114(526):668-681. [doi: 10.1080/01621459.2018.1429274]

Abbreviations

APACHE: Acute Physiology, Age, and Chronic Health Evaluation
AUC: area under the receiver operating characteristic curve
eICU: electronic intensive care unit
GLORE: Grid Binary LOgistic Regression
ICU: intensive care unit
MCCG: multicenter collaboration gateway
OR: odds ratio
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 07.06.20; peer-reviewed by R Duan, T Pereira, N Mohammad Gholi Mezerji; comments to author 21.09.20; revised version received 16.11.20; accepted 03.03.21; published 05.04.21.

Please cite as:
Park JA, Sung MD, Kim HH, Park YR
Weight-Based Framework for Predictive Modeling of Multiple Databases With Noniterative Communication Without Data Sharing: Privacy-Protecting Analytic Method for Multi-Institutional Studies
JMIR Med Inform 2021;9(4):e21043
URL: https://medinform.jmir.org/2021/4/e21043
doi:10.2196/21043
PMID:33818396

Original Paper

An Agent-Based Model of the Local Spread of SARS-CoV-2: Modeling Study

Alessio Staffini[1,2,3], MSc; Akiko Kishi Svensson[3,4,5], MD, PhD; Ung-Il Chung[3,6,7], MD, PhD; Thomas Svensson[3,4,6], MD, PhD

[1]Department of Economics and Finance, Catholic University of Milan, Milan, Italy

[2]Project Promotion Department, ALBERT Inc, Tokyo, Japan

[3]Precision Health, Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Tokyo, Japan

[4]Department of Clinical Sciences, Lund University, Malmö, Sweden

[5]Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

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

[7]Clinical Biotechnology, Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan

Corresponding Author:
Alessio Staffini, MSc
Precision Health, Department of Bioengineering
Graduate School of Engineering
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku
Tokyo, 113-8655
Japan
Phone: 81 080 7058 1309
Email: alessio.staffini@bocconialumni.it

Abstract

Background: The spread of SARS-CoV-2, originating in Wuhan, China, was classified as a pandemic by the World Health Organization on March 11, 2020. The governments of affected countries have implemented various measures to limit the spread of the virus. The starting point of this paper is the different government approaches, in terms of promulgating new legislative regulations to limit the virus diffusion and to contain negative effects on the populations.

Objective: This paper aims to study how the spread of SARS-CoV-2 is linked to government policies and to analyze how different policies have produced different results on public health.

Methods: Considering the official data provided by 4 countries (Italy, Germany, Sweden, and Brazil) and from the measures implemented by each government, we built an agent-based model to study the effects that these measures will have over time on different variables such as the total number of COVID-19 cases, intensive care unit (ICU) bed occupancy rates, and recovery and case-fatality rates. The model we implemented provides the possibility of modifying some starting variables, and it was thus possible to study the effects that some policies (eg, keeping the national borders closed or increasing the ICU beds) would have had on the spread of the infection.

Results: The 4 considered countries have adopted different containment measures for COVID-19, and the forecasts provided by the model for the considered variables have given different results. Italy and Germany seem to be able to limit the spread of the infection and any eventual second wave, while Sweden and Brazil do not seem to have the situation under control. This situation is also reflected in the forecasts of pressure on the National Health Services, which see Sweden and Brazil with a high occupancy rate of ICU beds in the coming months, with a consequent high number of deaths.

Conclusions: In line with what we expected, the obtained results showed that the countries that have taken restrictive measures in terms of limiting the population mobility have managed more successfully than others to contain the spread of COVID-19. Moreover, the model demonstrated that herd immunity cannot be reached even in countries that have relied on a strategy without strict containment measures.

(JMIR Med Inform 2021;9(4):e24192) doi:10.2196/24192

KEYWORDS

computational epidemiology; COVID-19; SARS-CoV-2; agent-based modeling; public health; computational models; modeling; agent; spread; computation; epidemiology; policy

Introduction

The spread of communicable diseases across a population is a spatial and temporal process, and the study of the transmission dynamics is becoming increasingly important for tackling the spread appropriately.

Agent-based models (ABMs) are a class of computational models based on computer simulations of actions and interactions of autonomous agents, aimed at evaluating how these actions affect the system as a whole. The agent-based approach emphasizes the importance of learning through the agent-environment interaction. This approach is part of a recent trend in the computational models of learning toward developing new ways of studying autonomous organisms in virtual or real environments.

ABMs have proven particularly useful for answering public health–related questions that are typically unanswerable with the traditional epidemiological toolkit [1]. The use of ABMs for studying phenomena related to public health is not recent and has been used to study the spread of alcohol consumption [2] and eating disorders [3].

Agent-based simulation modeling has been used primarily in epidemiological studies of infectious diseases, including the study of the reactions of the immune system during an infection [4], the spread of malaria following the movement of mosquitoes in a village in Niger [5], and following the trend of the influenza virus [6]. Additionally, ABMs have been used to study the trend of chronic diseases [7] and to analyze the public health impact of influenza vaccinations in the United States and their cost-effectiveness, simulating scenarios where different age groups of the population were vaccinated [8].

More recently, ABMs have been used in population-based studies of COVID-19, in particular to analyze the effects of population characteristics [9,10] and of public health measures on the spread of SARS-CoV-2 [11,12]. The importance of ABMs in the face of a global pandemic is their ability to reproduce situations, starting from real data, otherwise not reproducible in reality.

In this study, we propose an epidemiological ABM for analyzing the propagation of an infectious disease in a network of human contacts; in particular, our model studies the effects of political decisions on the spread of SARS-CoV-2. Other works have been done studying this aspect [13-15], but the approach was to simulate different pre-established situations (eg, implementing containment measures or performing many diagnostic tests), evaluating their impacts. A work similar to our study [16] starts from the same research questions and arrives at similar conclusions but uses a completely different methodology. Our study differs from the previous ones in that it analyzes the effects of the measures adopted by the governments in *real time* as they are implemented. An increasingly used ABM for modeling COVID-19 is Covasim [17]; although we propose a similar model that includes demographic information and nonpharmaceutical interventions, we considered a simplified network structure (specifically, a dynamic random network, where the edges are created and destroyed at each period *t*) with a focus on capturing only the stylized facts for an immediate evaluation of the effects of changes in model parameters and in policies.

The time stamp (in days) of the model accurately reflects the timing of the political decisions taken from the end of January to July 1, 2020, and the model studies the evolution of the virus from its appearance up to a year later. The parameters we defined were derived from government policies, from real data provided by the government bodies, and from medical knowledge about the virus up to July 1, 2020; beyond this date, the model makes predictions of how the virus would have spread if all the considered variables would have followed the same evolution (for example, maintaining the containment measures as of July 1 in the 4 countries). There was no knowledge at the time about virus variants nor data about the vaccination campaign, so these have not been included.

The hypothesis from which we start is that the spread of a virus depends, in addition to epidemiological factors and the nature of the virus itself, on individual behavior or, more precisely, on political decisions that induce appropriate behavioral criteria. Our goal is to show how, through targeted measures, the damage caused by the spread of a pandemic can be limited, both in terms of the case-fatality rate and pressure on hospitals.

Methods

Overview of the Model

We implemented the model using NetLogo (free and open-source software, released under a GNU General Public License; Rel. 6.1.0), a multi-agent programmable modeling environment (source code available on GitHub [18]). The simulation was performed using the data of 4 countries (Italy, Germany, Sweden, and Brazil) that have had different policy approaches for the containment of SARS-CoV-2.

Italy was chosen in our analysis as it was the first country (after China) to report an important diffusion of the virus in its territory and had to make new decisions and implement measures without having the possibility to compare their effectiveness with those of other similar countries. Germany followed the example of Italy but with a much higher execution speed, relying also on a greater number of intensive care unit (ICU) beds (the highest in Europe; Source: National Center for Biotechnology Information [19]). Sweden took a different approach from other European countries: it did not deny the presence and the potential consequences of the virus spread in its territory but decided not to impose any limitation to individual freedoms, essentially aiming at obtaining herd immunity. Like Sweden, Brazil did not adopt national measures to contain the spread of the virus, despite the high number of deaths that this has caused.

A more detailed description of the differences and the reasons that led us to choose these 4 countries can be found in Multimedia Appendix 1.

The model studied, through the interactions between healthy individuals and infected individuals, how the virus spread over time and how the actions implemented by governments influenced its propagation. Starting from objective data provided by government bodies (Table 1), key variables related to the country, population, virus, and implemented policies were taken into consideration. We provide a complete list of the measured variables (Table 2; several indexes are the proportional transformation of the values obtained from Table 1, defined in the calibration phase of the model). We measured and demonstrated the results of how these variables evolved over time.

Table 1. Reference data used for constructing the model.

Demographics	Data (%)	Notes
Italy		
Older than 65 years	22.60	Source: Eurostat [20]
Beds for seriously ill patients	0.26	Source: OECD[a] [21]
Recovery rate	76.89	Source: World Health Organization [22]
Case-fatality rate	14.55	Source: World Health Organization [22]
Seriously ill	2.40	Source: Ministero della Salute [23]
Hospitalization rate	25.40	Source: Ministero della Salute [23]
Not seriously ill	74.60	Source: Ministero della Salute [23]
Germany		
Older than 65 years	21.40	Source: Eurostat [20]
Beds for seriously ill patients	0.60	Source: OECD [24]
Recovery rate	91.15	Source: World Health Organization [25]
Case-fatality rate	4.67	Source: World Health Organization [25]
Seriously ill	1.48	Source: Worldometer [26]
Hospitalization rate	6.20	Source: Worldometer [26]
Not seriously ill	92.30	Source: Worldometer [26]
Sweden		
Older than 65 years	19.80	Source: Eurostat [20]
Beds for seriously ill patients	0.20	Source: OECD [27]
Recovery rate	12.74	Source: Worldometer [26]
Case-fatality rate	9.02	Source: World Health Organization [28]
Seriously ill	2.55	Source: Worldometer [26]
Hospitalization rate	25.68	Source: Worldometer [26]
Not seriously ill	69.30	Source: Worldometer [26]
Brazil		
Older than 65 years	8.60	Source: CIA[b] [29]
Beds for seriously ill patients	0.19	Source: AMIB[c] [30]
Recovery rate	49.81	Source: World Health Organization [31]
Case-fatality rate	4.65	Source: World Health Organization [31]
Seriously ill	2.00	Source: Worldometer [26]
Hospitalization rate	8.00	Source: Worldometer [26]
Not seriously ill	90.00	Source: Worldometer [26]

[a]OECD: Organisation for Economic Co-operation and Development.

[b]CIA: Central Intelligence Agency.

[c]AMIB: Associação Medicina Intensiva Brasileira.

Table 2. List of the model variables.

Variables	Italy	Germany	Sweden	Brazil
Total population (units)	1000	1000	1000	1000
Older than 65 years (units)	230	210	200	90
Initial infectious people (units)	6	6	6	6
Transmissibility rate (%)	0.30	0.30	0.30	0.30
Immunity duration: mild cases (days)	100	100	100	100
Immunity duration: severe cases	Lifetime	Lifetime	Lifetime	Lifetime
Initial productivity index	2.0	2.0	2.0	2.0
Noncontagion index	0.01	0.01	0.01	0.01
Virus recognition	After 100 cases	After 60 cases	After 100 cases	After 100 cases
Beds for seriously ill patients (units)	13	30	20	9
Recovery index	5.0	6.0	1.5	2.5
Case-fatality index	0.8	0.3	0.4	0.3
Seriously ill index	1.5	1.0	1.5	1.3
Not seriously ill index	5.0	5.0	4.7	6.0
Mask use (decrease in transmissibility; %)	14.3	14.3	0	7.15
Physical distancing (decrease in transmissibility; %)	10.2	10.2	10.2	5.1
Infected tourists (max number; units)	2	2	1	1

Description of the Model

The model examined a sample of the population of each of the 4 countries, fixed at 1000 ($i \in \{1,2,...,1000\}$). It was similar to a small neighborhood of a city where the characteristics of the entire population are reproduced (the data we were interested in are reported in Tables 1 and 2). We assumed that an outbreak of COVID-19 has developed in this neighborhood. Naturally, government provisions were applied to this neighborhood as they were issued and with the same timing.

To better explain the logic behind our choices, we should imagine the considered space where the agents live as a laboratory where we applied the different policies, compliance to nonpharmaceutical measures, health structures, knowledge about the virus, etc. The *laboratory* space and the number of agents go beyond the geographical context, as they are meant to represent an exportable sample for each of the analyzed countries. Only the different measures and country specifications influence the results obtained from the simulations.

The time span of the simulation was 1 year, divided into 365 daily cycles. The first cycle coincided with the first infections in the given country.

The initialization of the model (at time t=0) requires the loading and setting of the required variables for the simulations and analyses. The variables derived from the national and governmental bodies, as well as institutional sources for each country, were automatically loaded following country selection. In this phase, the model also set the time stamps in which political decisions were made with respect to the containment measures for the spread of SARS-CoV-2. After the initial setting, the model was ready to simulate the evolution of COVID-19 for the selected country.

The following shows the (simplified) scheme followed by the model. The total 1000 agents move randomly within the model environment, simulating daily activities (eg, going to work or school); movement speed was set at a lower level (50%) for older adult agents (older than 65 years), as they perform fewer activities (the number of older adults was modeled according to the national statistics, see Table 2). Among the agents, some are infected (we denoted them as I_i, and we fixed them at 6 at time t=0). Moving inside the environment, they come into contact with healthy agents. A healthy agent H_i has a certain probability $P(I) \in [0,1]$ to be infected, defined by the following equation:

$$P(I) = 1 - (1 - TR)^n$$

where $TR \in [0,1]$ is the transmissibility rate of the virus, and n is the number of infected neighbors; in our topology the number of infected agents present in a closed ball was determined by $B_1 = \{x \in R^2: \| \text{x-y} \| \leq 1\}$, with radius 1 and center $y \in R^2$, where the healthy agent H_i is located in time t. Notice that $P(I)$ is monotonically increasing with respect to TR and n.

The propagation of the virus is not immediately recognized as such by the governments, and before this happens, the number of infected agents I_i exceeds a certain threshold (see *virus recognition* in Table 2 for the country-specific threshold values).

After the virus is recognized, at each time t, we assume that each infected agent and those who have come into contact with them have a probability $P(test)$=0.5 to perform a virus recognition test [32]. Therefore, half of them do not perform

the test and continue to move inside the model space becoming, if infected, a symptomatic infected agent or an asymptomatic infected one. Asymptomatic agents will perform the virus recognition test at the next period $t + 1$ only if they come into contact again with an infected agent, while symptomatic agents will have the same constant probability $P(test)$ to be tested in each period.

Infected agents do not present symptoms immediately, but we considered that there is an incubation period that can vary according to age. For individuals younger than 65 years, we set the incubation period according to a normal distribution with mean 7 (SD 2; $I_Y \sim N(7,4)$), while for those who are older than 65 years, the incubation period is defined according to a normal distribution with mean 3 (SD 1; $I_O \sim N(3,1)$).

The viral load, and therefore the ability to infect other agents, has not been set the same for all of the I_i agents. For those who are in the incubation phase, the viral load is lower, and it increases period by period as the development of the infection approaches; for asymptomatic agents, it is lower than for agents with mild symptoms, who in turn, will have it lower than seriously ill agents (that will need to be hospitalized).

Infected agents I_i can therefore be of four types: in incubation, asymptomatic, mildly ill, and seriously ill (see Tables 1 and 2). The mildly ill, when found, are isolated at home; in our model, this translates to their mobility being set to 0 (but they can still spread the virus). The seriously ill, when found, will be hospitalized, and their mobility will also be set to 0. Furthermore, the latter are to be considered in an isolated space, so we also considered that they will not spread the virus anymore.

If ICU beds (see *beds for seriously ill patients* in Table 2) are saturated, seriously ill patients will be placed in home isolation (with mobility at 0), but their probability of recovery (see Tables 1 and 2) decreases.

Seriously ill patients can die with a probability equal to that listed in Table 1 (*case-fatality rate*) and Table 2 (*case-fatality index*). For older adults, this probability is higher (we consider their greater fragility and the possible presence of other existing pathologies). Dead agents are denoted with D_i; we set both their mobility and their transmissibility rate to 0.

Ill patients can recover with a probability equal to that listed in Table 1 (*recovery rate*) and Table 2 (*recovery index*).

Recovered agents develop antibodies to the virus (ie, they become immune). We denoted the immune agents with IM_i. For those who were seriously ill, we considered that their antibodies lasted for the whole simulation, while those who were mildly ill will develop an immunity that lasts only for 100 days [33] (see Table 2). Recent studies [34] confirm that there is a difference in the duration of immunity, which depends on the severity of the development of the disease.

We also considered noncontagious asymptomatic cases (ie, there is a small proportion of healthy individuals who, following infection, immediately develop antibodies without showing symptoms and never become carriers of the virus). They transition from H_i in t to IM_i in t +1 and are not counted as I_i. Notice that in each t, the sum of all the agents (healthy, infected, immune, and dead) is equal to 1000.

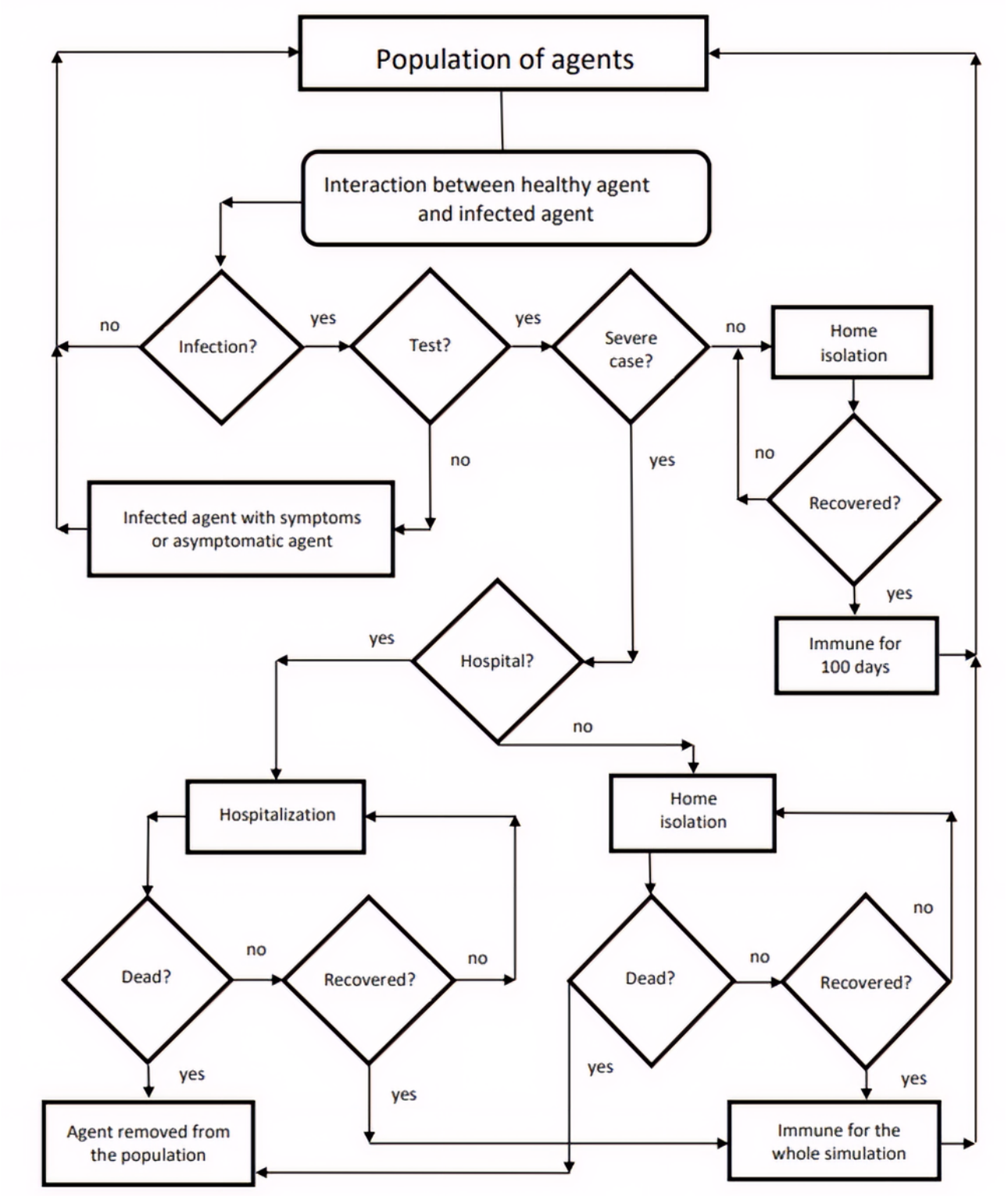
The industrial productivity (economic index) is proportional to the mobility of the agents. By setting the prepandemic level to 0, the reduced mobility of the agents will lead to a decrease in productivity.

Figure 1 shows a simplified flowchart of the mechanisms previously described.

Political decisions were then applied to this scheme, according to the times and the ways they have been implemented by the governments of the analyzed countries. Thus, for example, the decision of closing schools will lead to a reduction in the initial mobility of the agents; a lockdown of nonessential activities will further reduce it (with negative repercussions on industrial productivity, but a positive result in terms of limiting the spread of the contagion). The adoption of precautions or medical aids was translated into the model as a decrease in the transmissibility rate (see Table 2). We analyzed only a small sample of the population; therefore, in the rest of this study, the political decision of closing the national borders was translated into a further limitation on the mobility of agents, while their reopening was simulated as a partial restoration of the original mobility and the introduction of new infected agents (see *infected tourists* in Table 2).

A more detailed explanation of the parameters and the variables we used can be found in Multimedia Appendix 1.

Figure 1. Simplified flowchart of interaction mechanisms in the model.



Results

It should be remembered that the reported results were obtained considering the government measures in force until July 1, 2020; from this date onward, the forecasts are based on the last known measures being kept in place. In the summer of 2020, individual behavior and governments' attitudes were not so strict: therefore, despite the model correctly forecasting a second wave, such forecasts were underestimated.

Italy

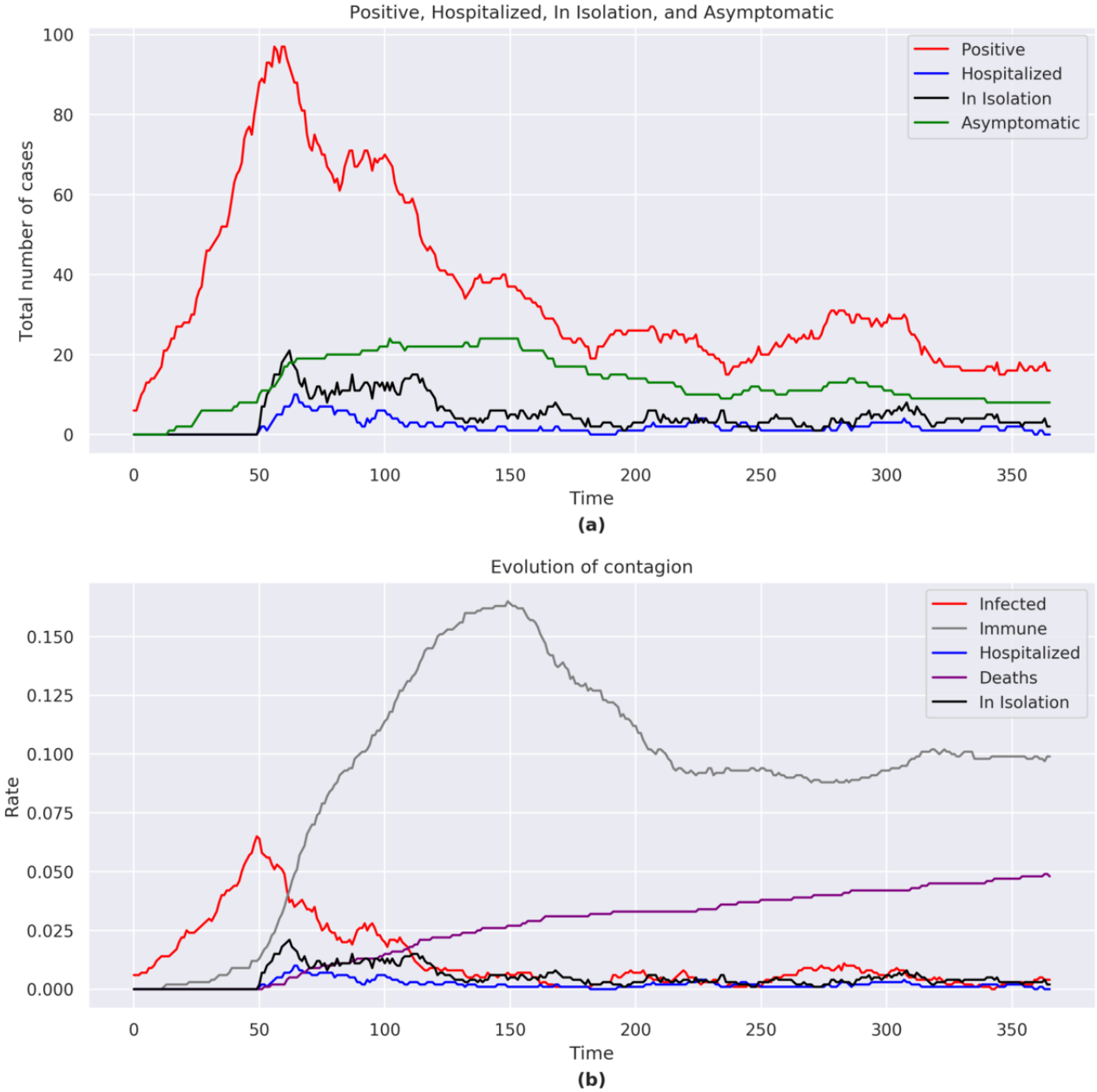
The simulation recorded 309 cases of COVID-19 with 243 recoveries and 48 deaths. The case-fatality rate was 15% (48/309), with an older adult (older than 65 years) case-fatality rate of 65% (31/48).

Total Number of COVID-19 Cases, Hospitalizations, Isolations, and Asymptomatic Infections

At the end of the simulation, the total number of positives was 17, including 8 asymptomatic, 2 in home isolation, and 0 hospitalizations. The number of COVID-19 cases rose exponentially with a peak at *t*=61 (Figure 2a). In the second half of the simulation (*t*=279), the number of COVID-19 cases rose but never reached the height of the initial peak.

The total number of hospitalizations and home isolations reached a peak at *t*=64.

Figure 2. (a) Positive, hospitalized, in isolation, and asymptomatic figures for Italy. (b) Evolution of the contagion for Italy. The graphs consider the sum of the agents belonging to each category shown in the legend for each day of the simulation.



Immunity and Case Fatality

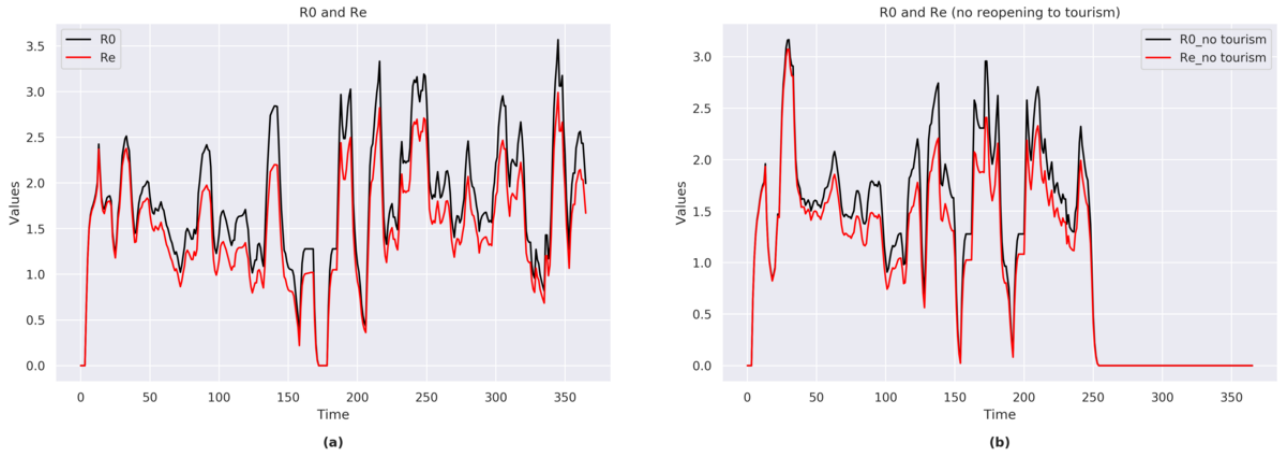
The proportion of individuals who acquired immunity reached a peak of 16.5% (165/1000) at *t*=150, albeit with an immunity of 10% (99/1000) at the end of the simulation (Figure 2b). The case-fatality rates increased throughout the simulation despite a decreasing number of cases. At the end of the simulation, the case-fatality rate was 4.8% (48/1000).

R_0 and R_E

The trends of R_0 (range 0-3.5) and R_E (range 0-3.0) exhibited strong fluctuations during the time span of the simulation (Figure 3a).

A sensitivity analysis conducted in a simulation that assumed that national borders remained closed (Figure 3b) showed that no new cases occurred once the *no contagion* value was reached, and the borders remained closed.

Figure 3. (a) R_0 and R_e indexes for Italy (with national borders reopening). (b) R_0 and R_e indexes for Italy (with no national borders reopening).



Herd Immunity, ICU Beds, and Productivity

The model showed that immunity was reached in approximately 11% (116/ 1000) of the population.

The simulation indicated that the ICU beds were never saturated; although at *t*=66, the occupancy rate reaches 77% (10/13).

The model showed a sharp drop in productivity following the implementation of containment measures, and the loss in productivity at its maximum reached –18.7% (compared to the prepandemic value of 0).

Germany

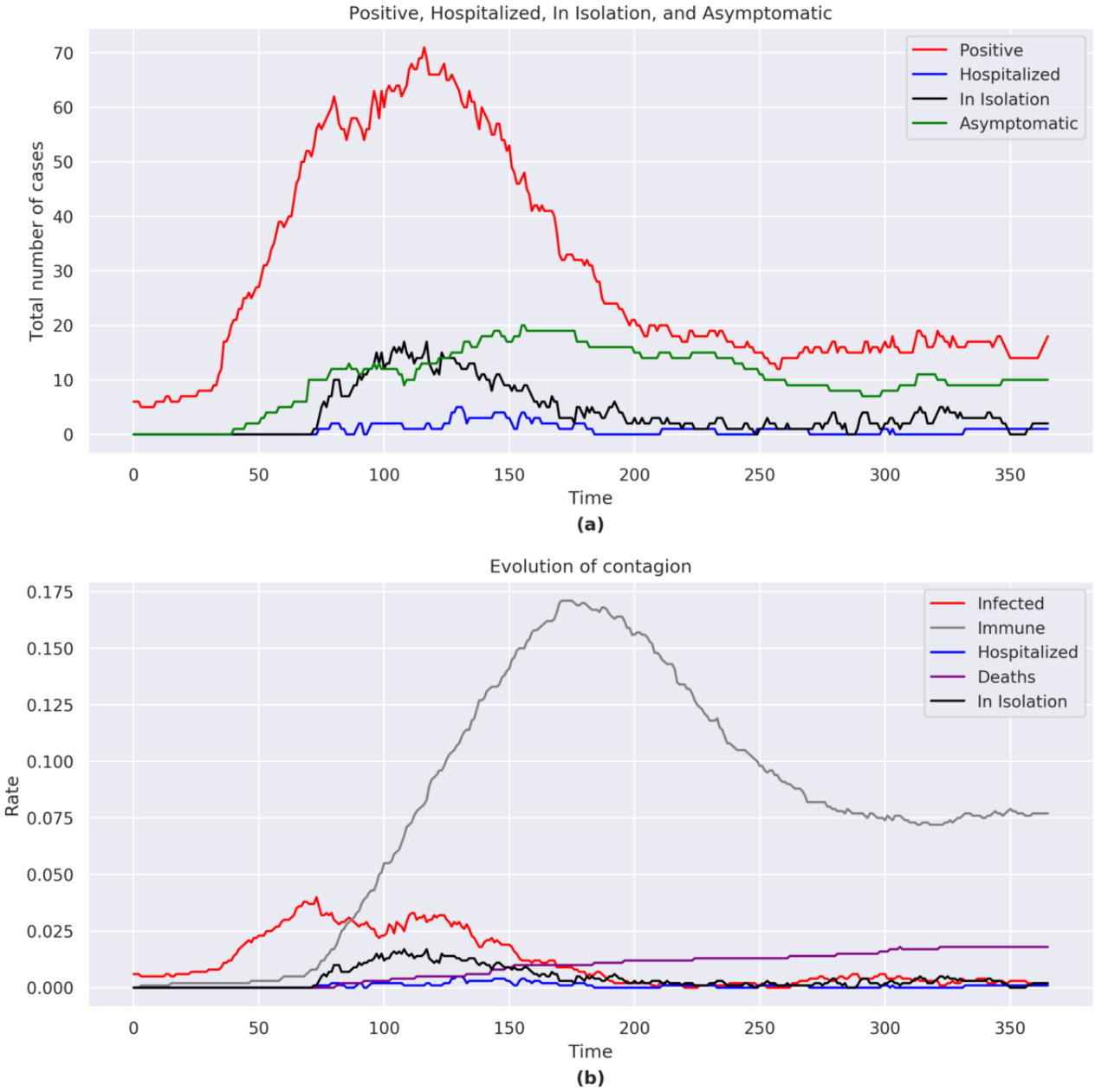
The simulation recorded 270 cases of COVID-19 with 233 recoveries and 18 deaths. The case-fatality rate was 6.7% (18/270), with an older adult case-fatality rate of 61% (11/18).

Total Number of COVID-19 Cases, Hospitalizations, Home Isolations, and Asymptomatic Infections

At the end of the simulation, the total number of positive cases was 19, including 10 asymptomatic, 2 in home isolation, and 1 hospitalized. The number of COVID-19 cases rose rapidly with a peak at *t*=117 (Figure 4a). After the peak and following the adopted containment measures, the number of COVID-19 cases gradually decreased.

The total number of hospitalizations and home isolations reached a peak around *t*=110.

Figure 4. (a) Positive, hospitalized, in isolation, and asymptomatic figures for Germany. (b) Evolution of the contagion for Germany. The graphs consider the sum of the agents belonging to each category shown in the legend for each day of the simulation.



Immunity and Case Fatality

The proportion of individuals who acquired immunity reached a peak of 17.1% (171/ 1000) at *t*=175, albeit with an immunity below 8% (77/1000) at the end of the simulation (Figure 4b). The case-fatality rate increased throughout the simulation. In the last part of the simulation, despite new cases of COVID-19, the case-fatality rate did not increase. At the end of the simulation, the case-fatality rate was 1.8% (18/1000).

R_0 and R_E

The trends of R_0 (range 0-3.8) and R_E (range 0-3.0) exhibited strong fluctuations during the time span of the simulation.

The sensitivity analysis showed that no new cases occurred once the *no contagion* value was reached (same as the analysis for Italy; Figure 3b), and the borders remained closed. Conversely, in a situation with open national borders, COVID-19 was not completely eradicated despite the implemented measures.

Herd Immunity, ICU Beds, and Productivity

The model showed that immunity was reached in approximately 10% (96/1000) of the population.

The simulation indicated that the ICU beds were far from being saturated, with the highest rate being 17% (5/30) at *t*=131.

The model showed a sharp drop in productivity following the implementation of containment measures, with a maximum loss of productivity of –18.2% (compared to the prepandemic value of 0).

Sweden

The simulation recorded 765 cases of COVID-19 with 533 recoveries and 141 deaths. The case-fatality rate was 18.4% (141/765), with an older adult case-fatality rate of 43% (61/141).

Total Number of COVID-19 Cases, Hospitalizations, Home Isolations, and Asymptomatic Infections

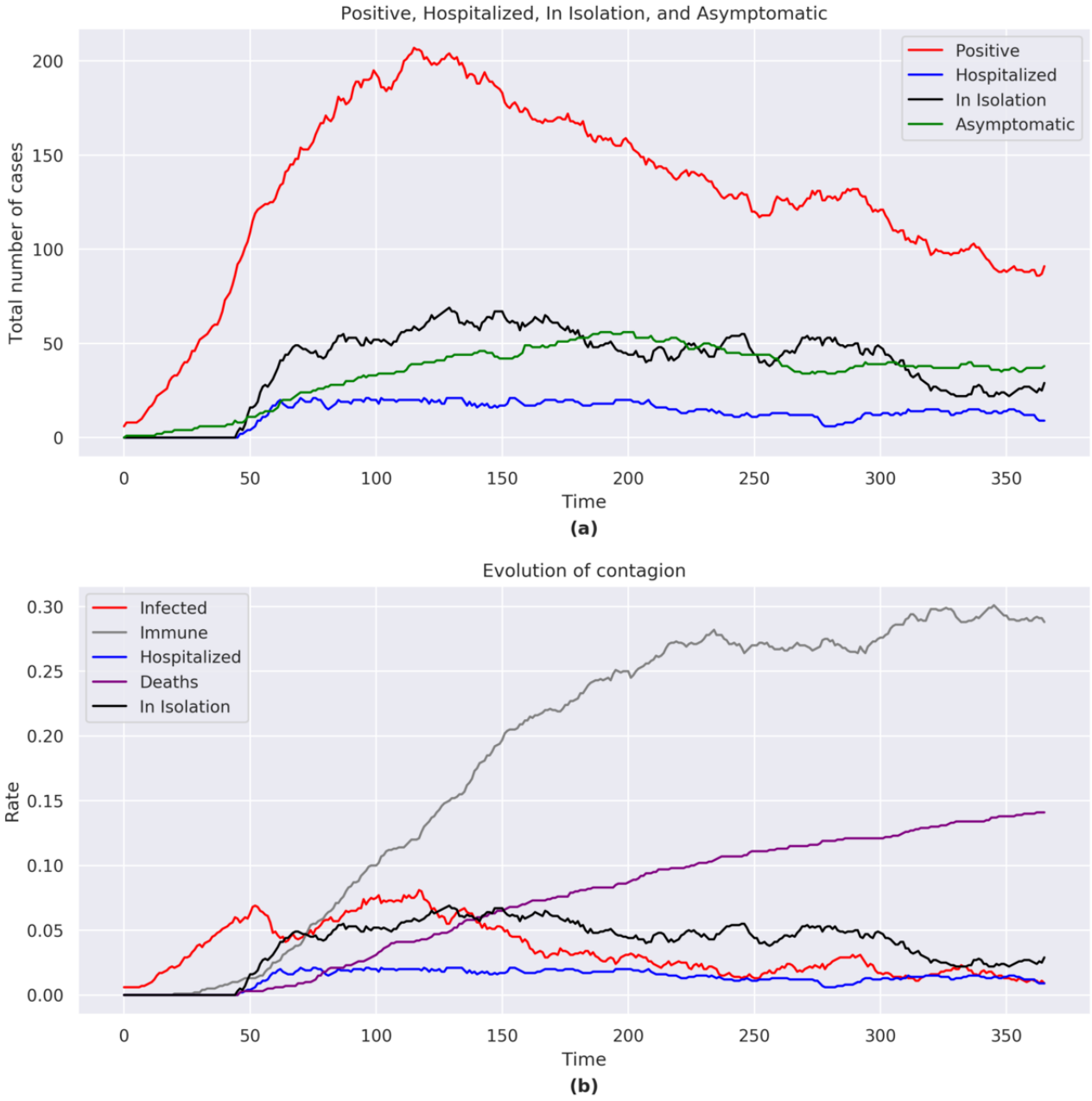
At the end of the simulation, the total number of positive cases was 91, including 38 asymptomatic, 29 in home isolation, and 9 hospitalized.

The number of COVID-19 cases reached a peak at t=116 (Figure 5a) with the number of cases decreasing slowly, remaining at high values until the end of the simulation. Despite a descending trend in the second half of the simulation, there were situations where the number of cases increased again. Given the high number of positive cases, the infection was not considered under control.

The total number of hospitalizations and home isolations reached a peak around t=129.

A sensitivity analysis that did not place a limit on the number of ICU beds showed that in the second part of the simulation there was a sharper decrease, with an overall lower number of COVID-19 cases.

Figure 5. (a) Positive, hospitalized, in isolation, and asymptomatic figures for Sweden. (b) Evolution of the contagion for Sweden. The graphs consider the sum of the agents belonging to each category shown in the legend for each day of the simulation.



Immunity and Case Fatality

Immunity was reached in 29% (290/1000) of the population, with a COVID-19 mortality rate of 14% (141/1000) at the end of the simulation (Figure 5b). The recovery rate was 60% (443/733; we did not count the number of positive agents at the end of the simulation).

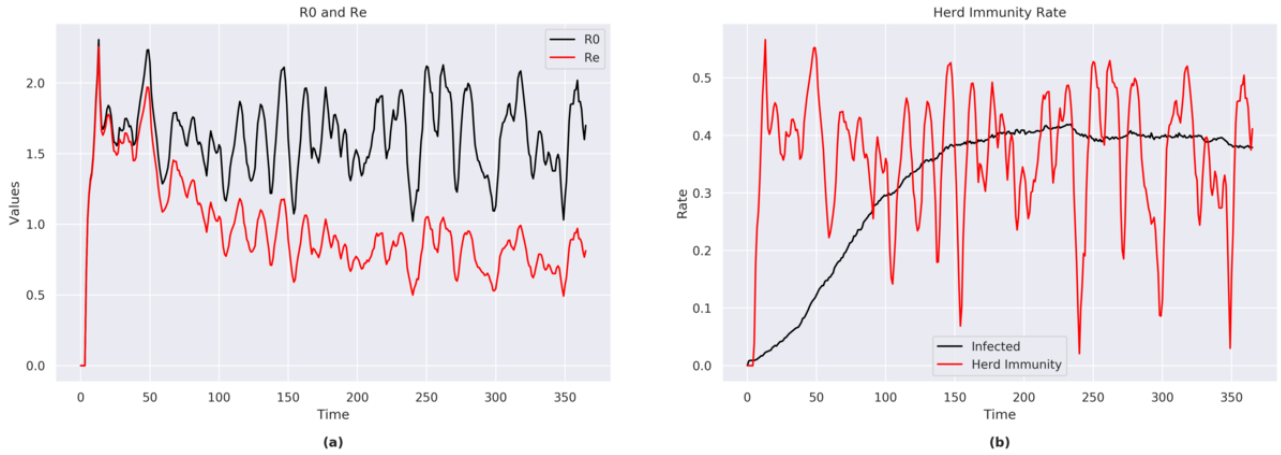
The sensitivity analysis showed that an increase in the number of ICU beds would lead to a decrease in the total number of positive cases and would result in an increased recovery rate (447/588, 76%; we did not count the number of positive agents at the end of the simulation) and a decreased case-fatality rate (80/693, 11.5%).

R_0, R_E, and Herd Immunity

The trends of R_0 (range 1.1-2.3) and R_E (range 0.5-2.2) presented less fluctuations during the time span of this simulation (Figure 6a).

The model showed that immunity was reached in approximately 39% (388/1000; we counted the immune plus the currently positive cases) of the population (Figure 6b).

Figure 6. (a) R_0 and R_e indexes for Sweden. (b) Herd immunity rate for Sweden.

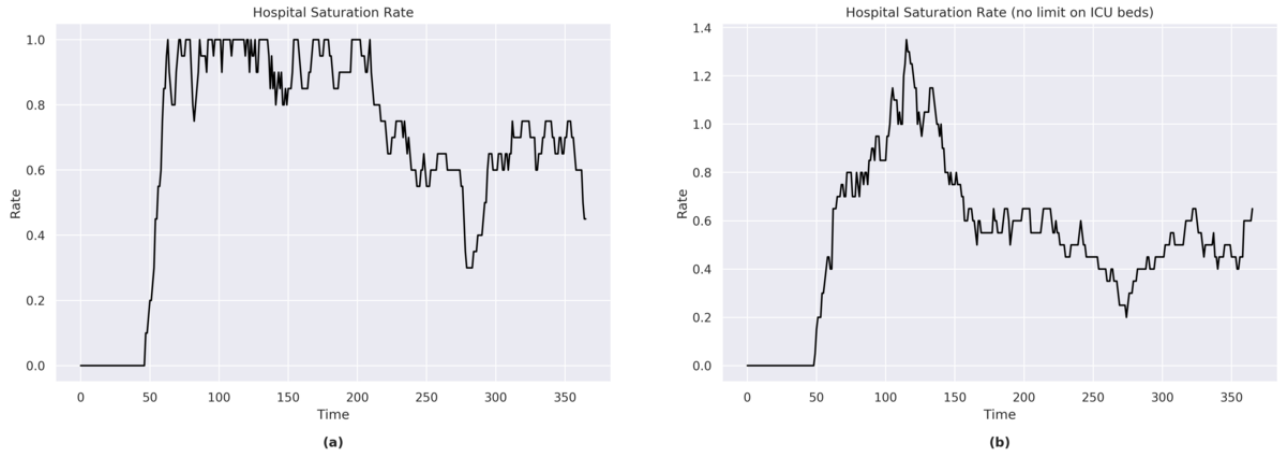


ICU Beds

In the first part of the simulation, full saturation of ICU beds was reached on a number of occasions (Figure 7a). In the second part of the simulation, the bed saturation rate remained above 50%.

The sensitivity analysis showed that approximately an additional 40% (an increase of 8 out of the current 20) of the available ICU beds would have been necessary to cope with the peak of a maximum emergency (at t=116; Figure 7b). An additional 10% of available beds would have met the needs of the population throughout most of the simulation period.

Figure 7. (a) Hospital saturation rate for Sweden. (b) Hospital saturation rate for Sweden (with no limit on the number of ICU beds). ICU: intensive care unit.



Productivity

The loss in productivity at its maximum reached –18.8% (compared to the prepandemic value of 0). The government measures have generated a decrease in productivity, and as these measures were still in place as of July 1, 2020, we could not see a rise due to the restoration of normality. The loss of productivity was mainly affected by the high number of infected individuals in hospital and in home isolation, and the high case-fatality rate.

In the sensitivity analysis without a limit on ICU beds, the productivity was higher than in the main analysis.

Brazil

The simulation recorded 883 cases of COVID-19 with 658 recoveries and 90 deaths. The case-fatality rate was just over 10% (90/883), with an older adult case-fatality rate of 34% (31/90) out of total deaths.

Total Number of COVID-19 Cases, Hospitalizations, Home Isolations, and Asymptomatic Infections

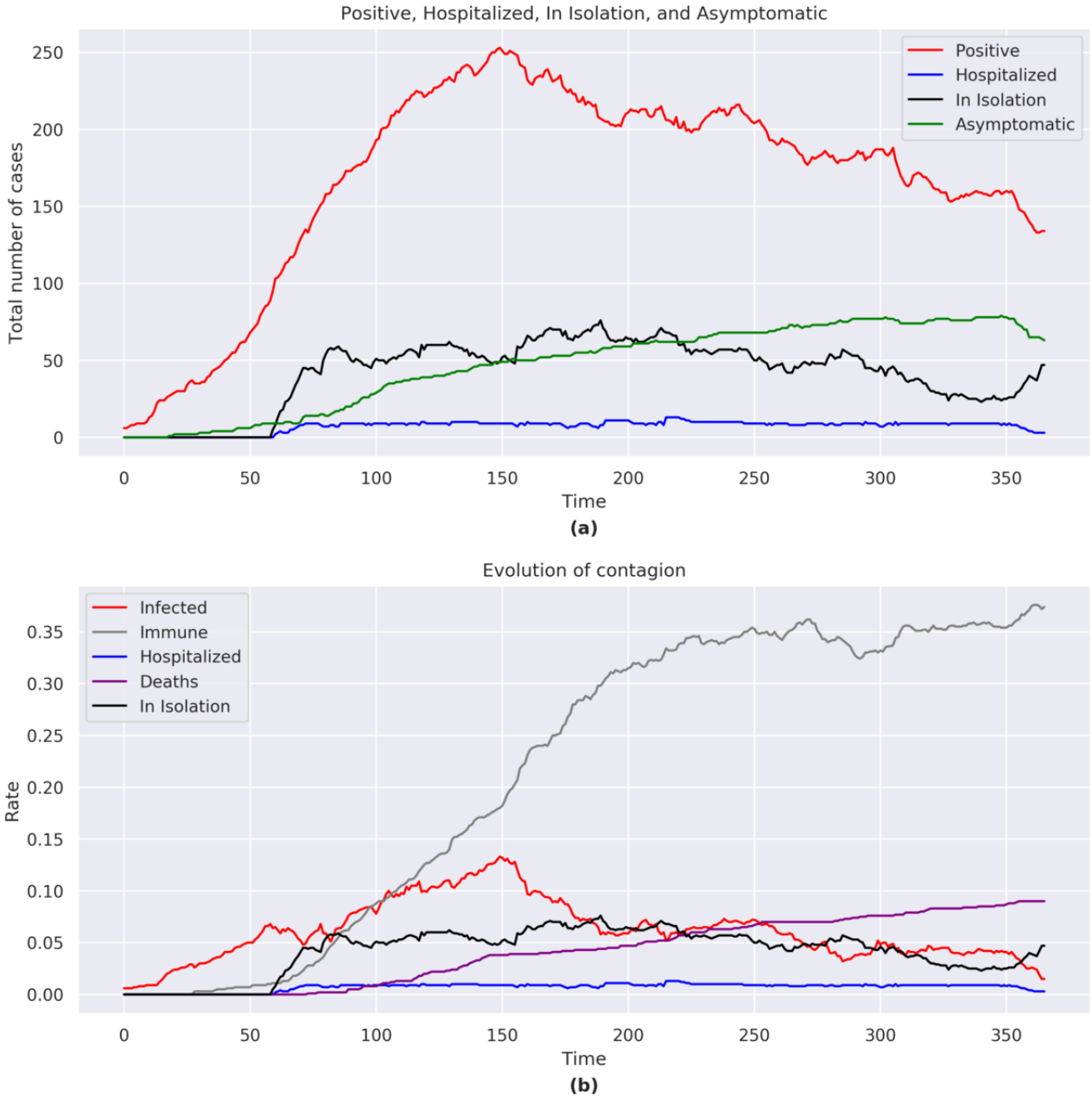
At the end of the simulation, the total number of positive cases was 137, including 63 asymptomatic, 47 in home isolation, and 3 hospitalized.

The number of COVID-19 cases gradually rose throughout most of the simulation period without reaching a clear peak. The final part of the simulation showed that there was a decrease in the number of cases due to the large number of immune individuals. Given the high number of people still positive at the end of the simulation, the infection was not to be considered under control (Figure 8a).

The total number of hospitalizations and home isolations reached a peak around t=190.

A sensitivity analysis that did not place a limit on the number of ICU beds resulted in a peak of the number of COVID-19 cases at t=75, followed by a gradual decrease in the number of cases.

Figure 8. (a) Positive, hospitalized, in isolation, and asymptomatic figures for Brazil. (b) Evolution of the contagion for Brazil. The graphs consider the sum of the agents belonging to each category shown in the legend for each day of the simulation.



Immunity and Case Fatality

The proportion of immune individuals reached 37% (374/1000) of the population (Figure 8b). The proportion of recovered cases reached 74% (658/883).

The sensibility analysis showed that an increase in the number of ICU beds would lead to a decrease in the number of cases and would result in a recovery rate of 84% (415/492) and a case-fatality rate of 7% (34/492).

R_0 and R_E, Herd Immunity, ICU Beds, and Productivity

The trends of R_0 (range 0.5-4.1) and R_E (range 0.2-2.5) exhibited contained fluctuations during the time span of this simulation but always remained higher than 1.

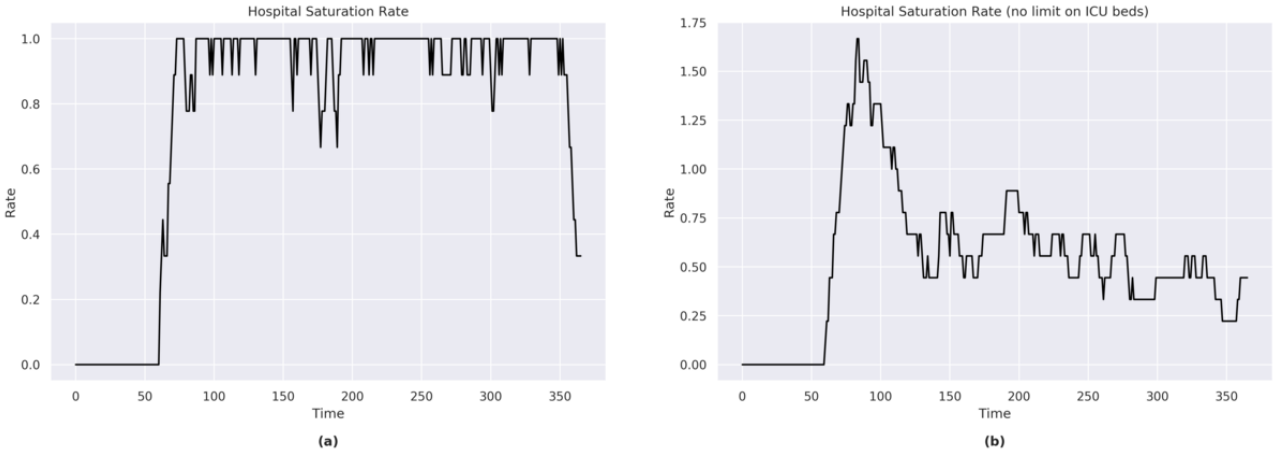
The model showed that immunity was reached in approximately 51% (511/1000) of the population, while a figure of 75% was required to obtain herd immunity for the entire population.

The simulation model showed that the number of ICU beds were insufficient with respect to the needs resulting from the spread of the COVID-19 pandemic (Figure 9a).

The sensitivity analysis, which removed the limit on the number of ICU beds, showed that, approximately, an additional 75% (an increase of 7 out of the current 9) of the number of available ICU beds would have been necessary to cope with the peak of maximum emergency (at t=84; Figure 9b).

The loss in productivity at most reached –12.4% (compared to the prepandemic value of 0).

Figure 9. (a) Hospital saturation rate for Brazil. (b) Hospital saturation rate for Brazil (with no limit on the number of ICU beds). ICU: intensive care unit.



Discussion

Objective

The objective of this study was to present a model that simulates the propagation of the COVID-19 pandemic based on real-world containment measures, as they were implemented by the governments of 4 countries: Italy, Germany, Sweden, and Brazil. The model thus allows for a prediction on the evolution of COVID-19 by reporting forecasts on key indexes such as the case-fatality rate, the recovery rate, herd immunity, ICU bed occupancy rates, home isolation rates, and the countries' productivity rates. The proposed model is highly flexible and allows for the addition or removal of parameters such as requirements and policies. Moreover, the model consequently studies how the contagion evolves over time. This offers the possibility to run additional simulations that predict the course of the pandemic under alternative policies by each government.

Previous models of SARS-CoV-2 have assessed the impact of the use of personal protection and early diagnosis [11], studied the impact of face masks on the spread of the virus [12], and analyzed the impact of the virus according to age [9,10], family situation, and the presence of comorbidities [10]. Meanwhile, other ABMs have considered the impact of home isolation on the saturation of ICU beds [35], assessed infection and fatality rates assuming a 20-fold underreported number of cases [36], or hypothesized the economic effects in Japan of a Tokyo lockdown [37]. Our model thus differs from previous models, as it focuses on the effects of contagion and on its evolution over time, considering both the real data made available by government bodies and the policy measures implemented to stop or limit the propagation of SARS-CoV-2.

The model outputs shown in this paper are the results of several simulations for each country. Due to the nature of ABMs, the quantitative results will differ with each simulation. Any conditions that occur within the model will vary over time while maintaining parameter values and keeping initial variables constant. Although each simulation will not yield identical quantitative results for each country, the qualitative behavior always follows the same trend. Consequently, we have been able to draw some considerations about the analyzed parameters. These are presented on a country by country basis.

Italy

The simulations for Italy show a low total number of COVID-19 cases compared to the simulations for Sweden and Brazil, indicating a success of the adopted containment measures. Similarly, the numbers of hospitalized individuals and those in home isolation seemed to remain under control. Overall, this resulted in a large fluctuation of R_0 and R_E, where a small increase in the number of infections lead to a large growth in the indexes' values. Additionally, the simulations for Italy indicated a slow reduction in the number of asymptomatic

individuals, which highlights an increased possibility of new infections that in turn could be extended to a recommendation not to loosen the implemented containment measures as of July 1, 2020. This is further supported by the trends in immunity, where the proportion of immune individuals were comparatively low. Bearing in mind that our model was implemented at the end of June 2020 and that it used statistical data available at that time, it was able to correctly predict that, with the reopening of the national borders and the free movement of people, there would be a new increase in the number of positive cases (thereby partly invalidating national containment efforts). This was indeed confirmed in our sensitivity analysis where the national borders remain closed. Such a situation is not replicable in reality, but it leads to no new cases.

Despite its relatively low number of cases, Italy recorded the second highest case-fatality rate (48/309, 15%; Germany: 18/270, 6.7%; Sweden: 141/765, 18.4%; Brazil: 90/883, 10%). Italy's proportion of older adults ranks among the highest in the world (Source: Istituto Nazionale di Statistica [38]), which could serve to partially explain the exceedingly high case-fatality rate. Indeed, the simulation indicates that COVID-19 affects older adults predominantly, where the case-fatality rate reached 65% (31/48) of total deaths.

Herd immunity in Italy would be obtained in a situation where 70% of the population are immune to SARS-CoV-2. The results of the model, however, indicated that only 11% of the Italian population reached immunity, a number that considers both immune individuals and active cases. This low proportion of immune individuals is expected given the policy decisions aimed at limiting the spread of the virus.

Germany

For Germany, our model was able to make a complete analysis of the contagion peak and its gradual descent, as well as predicted possible developments in the coming months.

The simulations for Germany showed a situation with a comparatively low number of infected individuals and strong fluctuations in R_0 and R_E indexes. The low infection rates resulted in a very low case-fatality rate; however, it also resulted in a low proportion of immune individuals. Like the simulations for Italy, the low number of positive cases and the low proportion of immune individuals was a consequence of the policy implementations aimed at containing the spread of SARS-CoV-2. Overall, the results at the end of the simulation for Germany are not too different from the results obtained for Italy, a situation under control with regard to hospitalizations and home isolation cases. Moreover, as for Italy, asymptomatic cases were still recorded (1%), which indicates a situation where SARS-CoV-2 is still present in the population, with the risk of continued virus spread if the containment measures were to be loosened. The model, starting from the data at the end of June 2020, correctly predicted that this percentage of asymptomatic people would have led to the formation of new outbreaks and a relatively new spread of the virus.

The simulations for Germany showed two notable differences compared to Italy. First, the proportion of recoveries was higher in Germany (233/251, 93% vs 243/292, 83%). Second, although the German simulations showed an older adult case-fatality rate of 61% (11/18), the overall case-fatality rate was only 6.7% (18/270). Germany's markedly lower overall case-fatality rate as compared to Italy could be the result of the prompt diagnosis and case management due to the widespread controls carried out by the public health authorities. The same containment measures, however, also result in the low proportion of immune individuals (96/1000, 10%), which are far from the proportion necessary to reach herd immunity as indicated by the model (73%). Consequently, and similar to Italy, the model predicted that Germany would have had a high risk of possible *second waves*, as it indeed happened with the reopening of the national borders.

Compared to Italy, Germany also fared better with regard to ICU bed occupancy rates; the simulations indicated that even in the most acute phase of the pandemic, bed occupancy rates never exceeded 20% (6/30) of total capacity. It should be kept in mind that Germany has by far the highest number of ICU beds in the 4 countries considered in our analysis [18].

Finally, the impact of the pandemic on the German economy is evident, as the containment measures had a strong impact on the productivity, which at one point reached –18.2%. However, contrary to the situation for Italy, the forecasts of major economic institutes such as the OECD and the World Bank considered the German recovery period to reach the precrisis values quicker than Italy.

Sweden

The simulations for Sweden demonstrate a situation that is not under control a year after the first recorded case and are thus in stark contrast to those obtained for Italy and Germany. These discrepancies are most likely due to the comparatively limited containment measures initiated by the Swedish Public Health Authority. First, both R_0 and R_E remain at higher values through the simulations, with an R_0 that never goes below 1. Second, at the end of the simulation, the total number of COVID-19 cases was much larger than in Italy and Germany; the high number of hospitalized and asymptomatic cases being of particular concern. Third, the proportion of recovered cases (443/733, 60%) was lower than the corresponding proportion in any of the other countries. Fourth, the case-fatality rate (141/765, 18.4%) was higher than the rates obtained for Germany, Italy, and Brazil.

A major difference of Sweden from Italy and Germany was the low number of available ICU beds. Despite its high focus on welfare, Sweden has a low number of ICU beds per capita. Although Sweden managed to double the number of ICU beds at the start of the pandemic (Source: Folkhälsomyndigheten [39]), the pressure on hospitals remains critical throughout the simulations. Indeed, as the sensitivity analyses showed, Sweden would have required an additional 40% of its ICU capacity at the peak of the pandemic. Moreover, the sensitivity analysis also showed that the case-fatality rate decreased from 18.4% to 11.5% with a higher number of ICU beds. It is therefore fair to conclude that an increase in the ICU capacity would have the potential to save many lives.

Despite the adverse outcomes, Sweden does not reach the threshold of herd immunity as determined by the simulations. The herd immunity threshold (57%) was derived based on specific considerations in the model (ie, lifelong immunity for those with serious COVID-19 and temporary immunity to those with milder forms of the disease). To the best of our knowledge, there is no clear data on immunity. Indeed, if the parameters in the model are accurate, herd immunity for COVID-19 would be difficult to reach. It is therefore possible to conclude, based on the simulations, that implementing containment measures and recommending the use of face masks have positive effects in limiting the spread and the consequences of COVID-19, even a year after the first recorded case.

The simulated productivity drop for Sweden would, unlike the situations in Italy and Germany, not be influenced by the country's containment measures but rather be a consequence of its large number of COVID-19 cases.

Brazil

As expected, our model foresees that Brazil has the highest number of COVID-19 cases among the 4 analyzed countries. This is most likely due to Brazil's implemented policy decisions, which are more in line with those of Sweden than Italy and Germany. Consequently, Brazil and Sweden share many similarities in the analyses. First, the total case numbers in Brazil resemble those of Sweden and are assumed to be a result of the less restrictive containments measures. Moreover, R_0 and R_E did not exhibit strong fluctuations, with R_0 remaining above 1 for the duration of the simulation. Despite a situation that could be considered out of control, Brazil displayed a notably lower case-fatality rate than Sweden (10.2% vs 18.4%). This is most likely due to the low proportion of older adults (8.6% vs 19.8% in Sweden). Another noteworthy difference between Brazil and Sweden was the encouraging recovery rate of 74% (658/883), again most likely due to the two countries' demographic differences in age.

The proportion of immune individuals in Brazil reached 37% (374/1000) of the population, the highest proportion of all the analyzed countries. Despite this high immunization rate, the model foresees that herd immunity will not be reached due to a calculated threshold of 75%. Indeed, the total proportion of immunized and positive cases at the end of the simulation reached 51% (511/1000).

The severity of the situation in Brazil was further highlighted by the sensitivity analyses, identifying a required 75% increase in the number of ICU beds for Brazil to cope with its situation. According to the models, such an increase in capacity would notably reduce not only the number of recorded COVID-19 cases but also the case-fatality rate (from 10.2% to 7%). It is, however, questionable whether an ICU capacity increase of such magnitude is feasible to implement, as Brazil over the past years has progressively decreased the availability of ICU beds (Source: Central Intelligence Agency [29]).

Brazil has adopted few measures concerning the closure of commercial activities (Source: Conselho Nacional de Secretários de Saúde [40]). This led to a lower drop in productivity (in absolute terms) than in European countries, with the difference that the contagion curve in Brazil lowers slowly; the model predicted that the negative effects of the pandemic will last for a long time so that it seems likely that other countries (that implemented stronger containment measures) will be able to reopen all their activities sooner. The productivity trend reflects this, as there is no rise toward pre–COVID-19 values.

The effects of reopening national borders cannot be assessed for Brazil, which unlike Germany and Italy, has never implemented closure of national borders as a measure to contain the spread of COVID-19.

Principal Results

By considering 4 countries with different policy approaches in the prevention and containment of the spread of COVID-19, our simulation model is able to highlight the consequences of policy decisions on a number of measures. The results obtained from our models showed the importance of prevention through widespread testing over large areas of territory (Germany) and of lockdown measures for the reduction of virus transmissibility (Italy and Germany). On the other hand, the countries that have not adopted these measures (Sweden and Brazil) are facing a situation that is not under control. From our results, we also highlight how important the mandatory use of face masks and the imposition of physical distancing are in reducing the number of COVID-19 cases. Our study also stresses how important it is to have an adequate number of ICU beds to deal with emergencies. This is evident particularly in the simulations for Sweden and Brazil, where the sensitivity analyses demonstrated an improvement in both recovery rates and case-fatality rates. Finally, the simulations showed that the reopening of national borders will not allow individual countries to maintain a monotonic decreasing curve of infections; indeed, only the simulations with the national borders being kept closed led to a complete stop of the spread of COVID-19.

In the context of an increasing number of positive COVID-19 cases, the main priority is the successful containment of the spread of SARS-CoV-2. However, prolonged lockdown measures have devastating effects on the economy of a country. The results of our model point toward a situation where countries that implemented mild policies against the virus at the start of the pandemic may inevitably need to strengthen them in the near future. Consequently, we suggest that the best course of action is to plan and implement aggressive political actions, both in the contagion containment phase (eg, limitations on the personal mobility and closure of nonessential activities) and in the economic recovery phase (eg, strong tax breaks for businesses and robust actions to stimulate consumption, as also indicated by the European Central Bank, even if doing this will result in a large budget deficit), with a long-term perspective from the beginning. According to the simulations, such actions may allow nations to overcome the economic impact of the pandemic sooner. This is important given that the data provided by the international economic organizations (International Monetary Fund, Organisation for Economic Co-operation and Development [OECD], World Bank, and others) leave no room for optimism [41-43].

Strengths and Limitations

There are a number of limitations that need to be mentioned. The main point concerns the input data for the model. We have retrieved the values from the most reliable sites among those that provide daily information about the spread of the virus, but this information is constantly evolving. Consequently, to keep the model updated, it is necessary to set up the most recent information. In this paper, the model *photographs* the situation at the end of June 2020, and it provides a forecast based on those data. Another limitation is that we considered only a small sample (which can be thought of as an infection outbreak). Even if this sample has the same national characteristics, the obtained results may not perfectly be the same when translated on a larger scale; that said, what we have obtained remains valid when studying a representative outbreak.

The economic results obtained from the model measured only the impact resulting from political decisions to contain the spread of COVID-19. The economic ramifications that will occur after a complete reopening of borders, such as a decrease in consumption and tourism, an increase in unemployment, and the shutdown of various economic activities, have not been taken into consideration.

The simulations also have a number of strengths. They take into consideration the age distribution of the respective countries. This is crucial given the impact of COVID-19 on the older adult population. The data in all the simulations is based on official statistics, as they are obtained through the national statistical databases of each country. This is a major strength for Sweden, Germany, and Italy, but a limitation for the analyses relating to Brazil. Moreover, the model can be extended to include additional new and relevant variables as they become available or are deemed necessary by researchers and policy makers.

Future Considerations

Further development of the model could allow for comparisons of the outcomes of a number of different policy proposals (eg, obligatory vs voluntary use of face masks, whether or not to increase the number of ICU beds, or whether or not to implement lockdown measures). The model could therefore be used to evaluate the needs and requirements for the considered territory, and the policies with the greatest impact over time. We plan to better explore these points in future research.

Additionally, with regard to the economic consequences of the pandemic, further considerations should be made for data concerning productivity and the economy in general. At the time of writing, the return to a situation similar to the one before the pandemic seems likely to occur only after the vaccination campaign ends, covering at least 75% of the population [44].

Conclusions

This study used real-world data to analyze how different political decisions aiming to deal with the spread of SARS-CoV-2 influence the extent of COVID-19. The results of the simulations lead to three main conclusions. First, strict containment measures, including the mandated use of face masks and the implementation of social distance, lead to a reduction in the number of COVID-19 cases. Second, the number of ICU beds are an important measure to reduce case-fatality rates. Third, herd immunity cannot be reached, and any national strategy aiming to reach herd immunity by loosening containment measures should be avoided.

Acknowledgments

The authors did not have any sources of funding to report. This research was supported by the Joint Research and Development Agreement between ALBERT Inc and the Center of Innovation at the University of Tokyo. This research was supported by the Center of Innovation Program of the Japan Science and Technology Agency (grant number JPMJCE1304).

Authors' Contributions

AS and TS developed and refined the ideas in this paper. AS wrote the first draft. AS and TS discussed it actively and revised the first draft. AKS and UIC critically revised the manuscript for important intellectual content, until the final agreement on the submitted version. All authors read and approved the final submitted version of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 27 KB - medinform_v9i4e24192_app1.docx]

References

1. Silverman E, Gostoli U, Picascia S, Almagor J, McCann M, Shaw R, et al. Situating agent-based modelling in population health research. arXiv 2021 Preprint posted online on February 6, 2020 [FREE Full text]
2. Gorman DM, Mezic J, Mezic I, Gruenewald PJ. Agent-based modeling of drinking behavior: a preliminary model and potential applications to theory and practice. Am J Public Health 2006 Nov;96(11):2055-2060. [doi: 10.2105/AJPH.2005.063289] [Medline: 17018835]

3. Zhang D, Giabbanelli PJ, Arah OA, Zimmerman FJ. Impact of different policies on unhealthy dietary behaviors in an urban adult population: an agent-based simulation model. Am J Public Health 2014 Jul;104(7):1217-1222. [doi: 10.2105/AJPH.2014.301934] [Medline: 24832414]
4. Chiacchio F, Pennisi M, Russo G, Motta S, Pappalardo F. Agent-based modeling of the immune system: NetLogo, a promising framework. Biomed Res Int 2014;2014:907171. [doi: 10.1155/2014/907171] [Medline: 24864263]
5. Bomblies A. Agent-based modeling of malaria vectors: the importance of spatial simulation. Parasit Vectors 2014 Jul 03;7:308 [FREE Full text] [doi: 10.1186/1756-3305-7-308] [Medline: 24992942]
6. Roche B, Drake JM, Rohani P. An agent-based model to study the epidemiological and evolutionary dynamics of Influenza viruses. BMC Bioinformatics 2011 Mar 30;12:87 [FREE Full text] [doi: 10.1186/1471-2105-12-87] [Medline: 21450071]
7. Li Y, Lawley MA, Siscovick DS, Zhang D, Pagán JA. Agent-based modeling of chronic diseases: a narrative review and future research directions. Prev Chronic Dis 2016 May 26;13:E69 [FREE Full text] [doi: 10.5888/pcd13.150561] [Medline: 27236380]
8. DePasse JV, Smith KJ, Raviotta JM, Shim E, Nowalk MP, Zimmerman RK, et al. Does choice of influenza vaccine type change disease burden and cost-effectiveness in the United States? An agent-based modeling study. Am J Epidemiol 2017 May 01;185(9):822-831 [FREE Full text] [doi: 10.1093/aje/kww229] [Medline: 28402385]
9. Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. Modelling transmission and control of the COVID-19 pandemic in Australia. Nat Commun 2020 Nov 11;11(1):5710. [doi: 10.1038/s41467-020-19393-6] [Medline: 33177507]
10. Wilder B, Charpignon M, Killian JA, Ou H, Mate A, Jabbari S, et al. Modeling between-population variation in COVID-19 dynamics in Hubei, Lombardy, and New York City. Proc Natl Acad Sci U S A 2020 Oct 13;117(41):25904-25910 [FREE Full text] [doi: 10.1073/pnas.2010651117] [Medline: 32973089]
11. Moore SE, Okyere E. Controlling the transmission dynamics of COVID-19. arXiv. Preprint posted online on March 31, 2020 [FREE Full text]
12. Kai D, Goldstein GP, Morgunov A, Nangalia V, Rotkirch A. Universal masking is urgent in the COVID-19 pandemic: SEIR and agent based models, empirical validation, policy recommendations. arXiv. Preprint posted online on April 22, 2020 [FREE Full text]
13. Louati D, Haddad G, Bedhiafi W, Bellamine N, Kebir A, Kchaou A, et al. ABM model to explore containment and screening policies to control COVID-19 virus spread. ResearchGate. Preprint posted online on March 1, 2020. [doi: 10.13140/RG.2.2.32409.16489]
14. Mahdizadeh Gharakhanlou N, Hooshangi N. Spatio-temporal simulation of the novel coronavirus (COVID-19) outbreak using the agent-based modeling approach (case study: Urmia, Iran). Inform Med Unlocked 2020;20:100403 [FREE Full text] [doi: 10.1016/j.imu.2020.100403] [Medline: 32835081]
15. Churches T, Jorm L. Flexible, freely available stochastic individual contact model for exploring COVID-19 intervention and control strategies: development and simulation. JMIR Public Health Surveill 2020 Sep 18;6(3):e18965 [FREE Full text] [doi: 10.2196/18965] [Medline: 32568729]
16. Kaxiras E, Neofotistos G. Multiple epidemic wave model of the COVID-19 pandemic: modeling study. J Med Internet Res 2020 Jul 30;22(7):e20912 [FREE Full text] [doi: 10.2196/20912] [Medline: 32692690]
17. Kerr CC, Stuart RM, Mistry D, Abeysuriya RG, Hart G, Rosefeld K, et al. Covasim: an agent-based model of COVID-19 dynamics and interventions. medRxiv. Preprint posted online on May 15, 2020. [doi: 10.1101/2020.05.10.20097469]
18. An Agent-Based Model of the Local Spread of SARS-CoV-2: Modeling Study. GitHub. URL: https://github.com/staale92/abm-local-pandemic-spread [accessed 2021-03-16]
19. National Center for Biotechnology Information. 2012. URL: https://www.ncbi.nlm.nih.gov/ [accessed 2020-06-01]
20. Eurostat. 2018. URL: https://ec.europa.eu/eurostat [accessed 2020-06-01]
21. Italy. OECD. 2018. URL: http://www.oecd.org/italy/ [accessed 2020-06-01]
22. Italy. World Health Organization. URL: https://www.who.int/countries/ita/ [accessed 2020-06-26]
23. Ministero della Salute. URL: http://www.salute.gov.it/portale/home.html [accessed 2020-06-01]
24. Germany. OECD. 2018. URL: http://www.oecd.org/germany/ [accessed 2020-06-01]
25. Germany. World Health Organization. URL: https://www.who.int/countries/deu/ [accessed 2020-06-26]
26. Worldometer. URL: https://www.worldometers.info [accessed 2020-06-01]
27. Sweden. OECD. URL: http://www.oecd.org/sweden/ [accessed 2020-06-01]
28. Sweden. World Health Organization. URL: https://www.who.int/countries/swe/ [accessed 2020-06-26]
29. Brazil. CIA. 2018. URL: https://www.cia.gov/the-world-factbook/countries/brazil [accessed 2020-06-01]
30. AMIB. 2018. URL: https://www.amib.org.br/ [accessed 2020-06-01]
31. Brazil. World Health Organization. URL: https://www.who.int/countries/bra/ [accessed 2020-06-26]
32. Böhning D, Rocchetti I, Maruotti A, Holling H. Estimating the undetected infections in the Covid-19 outbreak by harnessing capture-recapture methods. Int J Infect Dis 2020 Aug;97:197-201 [FREE Full text] [doi: 10.1016/j.ijid.2020.06.009] [Medline: 32534143]
33. Rodda LB, Netland J, Shehata L, Pruner KB, Morawski PA, Thouvenel CD, et al. Functional SARS-CoV-2-specific immune memory persists aftermMild COVID-19. Cell 2021 Jan 07;184(1):169-183.e17 [FREE Full text] [doi: 10.1016/j.cell.2020.11.029] [Medline: 33296701]

34. Ward H, Cooke G, Atchison C, Whitaker M, Elliott J, Moshe M, et al. Declining prevalence of antibody positivity to SARS-CoV-2: a community study of 365,000 adults. medRxiv. Preprint posted online on October 27, 2020. [doi: 10.1101/2020.10.26.20219725]
35. Shoukat A, Wells CR, Langley JM, Singer BH, Galvani AP, Moghadas SM. Projecting demand for critical care beds during COVID-19 outbreaks in Canada. CMAJ 2020 May 11;192(19):E489-E496 [FREE Full text] [doi: 10.1503/cmaj.200457] [Medline: 32269020]
36. Anastassopoulou C, Russo L, Tsakris A, Siettos C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. PLoS One 2020;15(3):e0230405 [FREE Full text] [doi: 10.1371/journal.pone.0230405] [Medline: 32231374]
37. Inoue H, Todo Y. The propagation of economic impacts through supply chains: the case of a mega-city lockdown to prevent the spread of COVID-19. PLoS One 2020;15(9):e0239251 [FREE Full text] [doi: 10.1371/journal.pone.0239251] [Medline: 32931506]
38. Istituto Nazionale di Statistica. URL: https://www.istat.it/ [accessed 2020-06-01]
39. Our mission – to strengthen and develop public health. Folkhälsomyndigheten. URL: https://www.folkhalsomyndigheten.se/the-public-health-agency-of-sweden/ [accessed 2020-06-01]
40. Conselho Nacional de Secretários de Saúde. URL: https://www.conass.org.br [accessed 2020-06-01]
41. Mishra MK. The world after COVID-19 and its impact on global economy. EconStar. 2020. URL: http://hdl.handle.net/10419/215931 [accessed 2020-06-30]
42. McKibbin W, Fernando R. The economic impact of COVID-19. In: Baldwin R, di Mauro BW, editors. Economics in the Time of COVID-19. London: Centre for Economic Policy Research; 2020.
43. Maliszewska M, Mattoo A, van der Mensbrugghe D. The potential impact of COVID-19 on GDP and trade: a preliminary assessment. World Bank Group. 2020. URL: https://openknowledge.worldbank.org/handle/10986/33605 [accessed 2020-06-30]
44. Kwok KO, Lai F, Wei WI, Wong SYS, Tang JWT. Herd immunity - estimating the level required to halt the COVID-19 epidemics in affected countries. J Infect 2020 Jun;80(6):e32-e33 [FREE Full text] [doi: 10.1016/j.jinf.2020.03.027] [Medline: 32209383]

Abbreviations

ABM: agent-based model
ICU: intensive care unit

Edited by C Lovis; submitted 08.09.20; peer-reviewed by A Chang, P Giabbanelli; comments to author 18.10.20; revised version received 06.11.20; accepted 21.03.21; published 06.04.21.

Please cite as:
Staffini A, Svensson AK, Chung UI, Svensson T
An Agent-Based Model of the Local Spread of SARS-CoV-2: Modeling Study
JMIR Med Inform 2021;9(4):e24192
URL: https://medinform.jmir.org/2021/4/e24192
doi:10.2196/24192
PMID:33750735

Original Paper

Automatable Distributed Regression Analysis of Vertically Partitioned Data Facilitated by PopMedNet: Feasibility and Enhancement Study

Qoua Her[1,2], PharmD, MSc, MSPharmD; Thomas Kent[3], PhD; Yuji Samizo[3], MS; Aleksandra Slavkovic[3], PhD; Yury Vilk[1,2], PhD; Sengwee Toh[1,2], ScD

[1]Department of Population Medicine, Harvard Medical School, Boston, MA, United States

[2]Harvard Pilgrim Health Care Institute, Boston, MA, United States

[3]Department of Statistics, Pennsylvania State University, University Park, PA, United States

Corresponding Author:
Qoua Her, PharmD, MSc, MSPharmD
Department of Population Medicine
Harvard Medical School
401 Park Drive, 4th Floor East
Boston, MA, 02215
United States
Phone: 1 6178674885
Email: qouaher@gmail.com

Abstract

Background: In clinical research, important variables may be collected from multiple data sources. Physical pooling of patient-level data from multiple sources often raises several challenges, including proper protection of patient privacy and proprietary interests. We previously developed an SAS-based package to perform distributed regression—a suite of privacy-protecting methods that perform multivariable-adjusted regression analysis using only summary-level information—with horizontally partitioned data, a setting where distinct cohorts of patients are available from different data sources. We integrated the package with PopMedNet, an open-source file transfer software, to facilitate secure file transfer between the analysis center and the data-contributing sites. The feasibility of using PopMedNet to facilitate distributed regression analysis (DRA) with vertically partitioned data, a setting where the data attributes from a cohort of patients are available from different data sources, was unknown.

Objective: The objective of the study was to describe the feasibility of using PopMedNet and enhancements to PopMedNet to facilitate automatable vertical DRA (vDRA) in real-world settings.

Methods: We gathered the statistical and informatic requirements of using PopMedNet to facilitate automatable vDRA. We enhanced PopMedNet based on these requirements to improve its technical capability to support vDRA.

Results: PopMedNet can enable automatable vDRA. We identified and implemented two enhancements to PopMedNet that improved its technical capability to perform automatable vDRA in real-world settings. The first was the ability to simultaneously upload and download multiple files, and the second was the ability to directly transfer summary-level information between the data-contributing sites without a third-party analysis center.

Conclusions: PopMedNet can be used to facilitate automatable vDRA to protect patient privacy and support clinical research in real-world settings.

(JMIR Med Inform 2021;9(4):e21459) doi:10.2196/21459

KEYWORDS

distributed regression analysis; distributed data networks; privacy-protecting analytics; vertically partitioned data; informatics; data networks; data

Introduction

Researchers often have to pool data from multiple sources for their studies. One common scenario is to combine data from multiple distinct cohorts of patients to achieve sufficient statistical power, especially in studies where the exposure or outcome of interest is rare. Another scenario is when important variables such as exposures, outcomes, or confounders are available from multiple data sources. However, physical pooling of patient-level data dispersed across multiple data sources often raises several concerns, including ownership of the data, unapproved use of the transferred data, and proper protection of patient privacy and proprietary interests of the data-contributing sites [1-5].

In most studies that analyze patient-level data from multiple sites, researchers can remove direct patient identifiers (eg, name, social security number) before sharing the data. It is also possible to relativize certain data attributes such as dates (eg, by setting the cohort entry date as time zero and converting all dates to numerical values relative to the time zero), perturb data attributes that may be used to reidentify patients (eg, rare covariates or laboratory values), or encrypt the deidentified patient level. However, these data manipulation techniques may not be feasible in certain studies and do not always guarantee adequate levels of privacy protection, which may deter collaboration and data sharing.

A number of privacy-protecting analytic methods have been developed to complement available data manipulation techniques. These methods, including meta-analysis of site-specific effect estimates and methods that leverage confounder summary scores, generally only require data-contributing sites to share summary-level information, thereby offering better privacy protection [6-8]. However, most methods were developed to analyze horizontally partitioned data (Figure 1), a setting where distinct cohorts of patients with the same data attributes are available in multiple data-contributing sites [9,10]. There are few valid and practical privacy-protecting methods to analyze vertically partitioned data (Figure 2), a setting where the data attributes of one distinct cohort of patients are available in two or more data-contributing sites.

Figure 1. Distributed regression analysis in horizontally partitioned data environments. In this hypothetical example, surgery, sex, and race are the independent variables, while BMI is the dependent variable. Both data-contributing sites have the same set of variables.

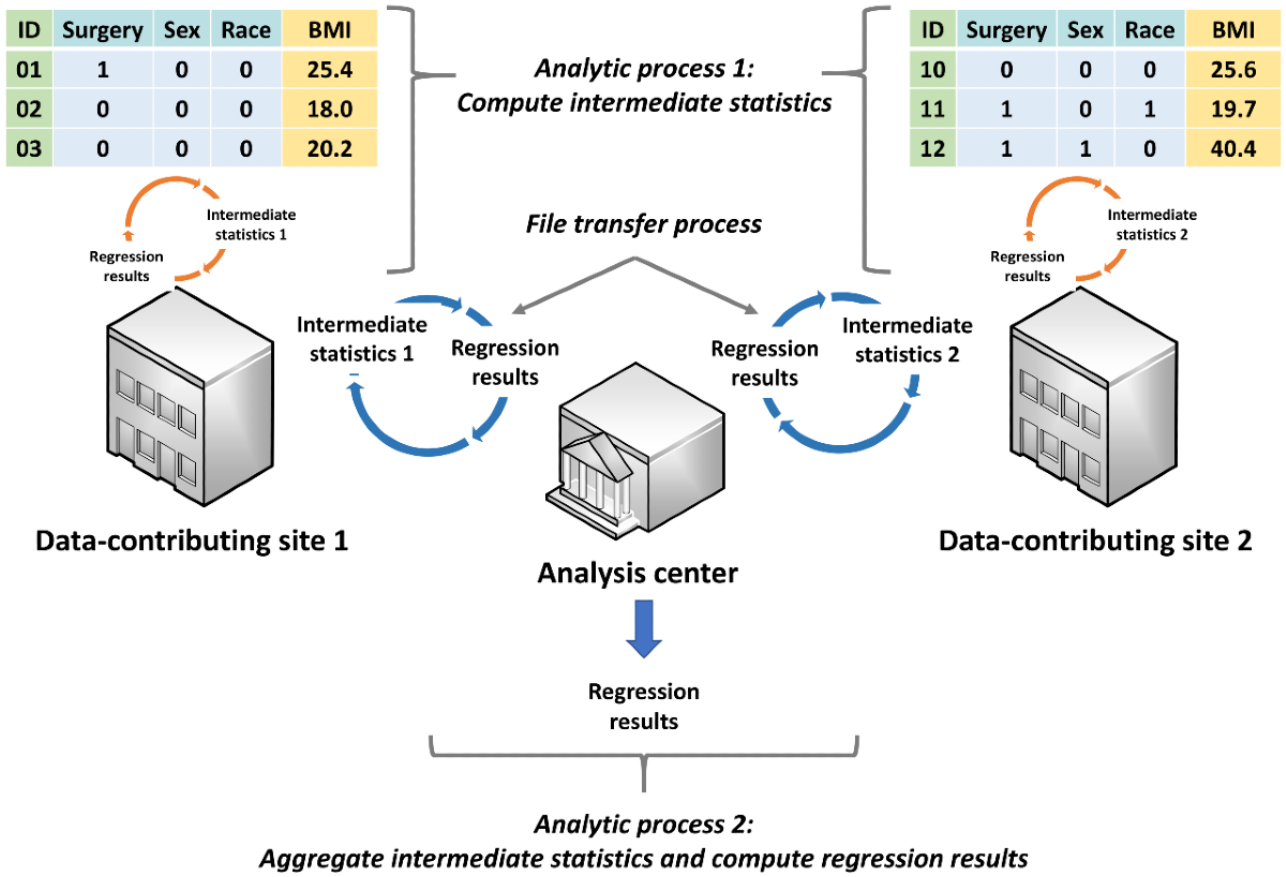
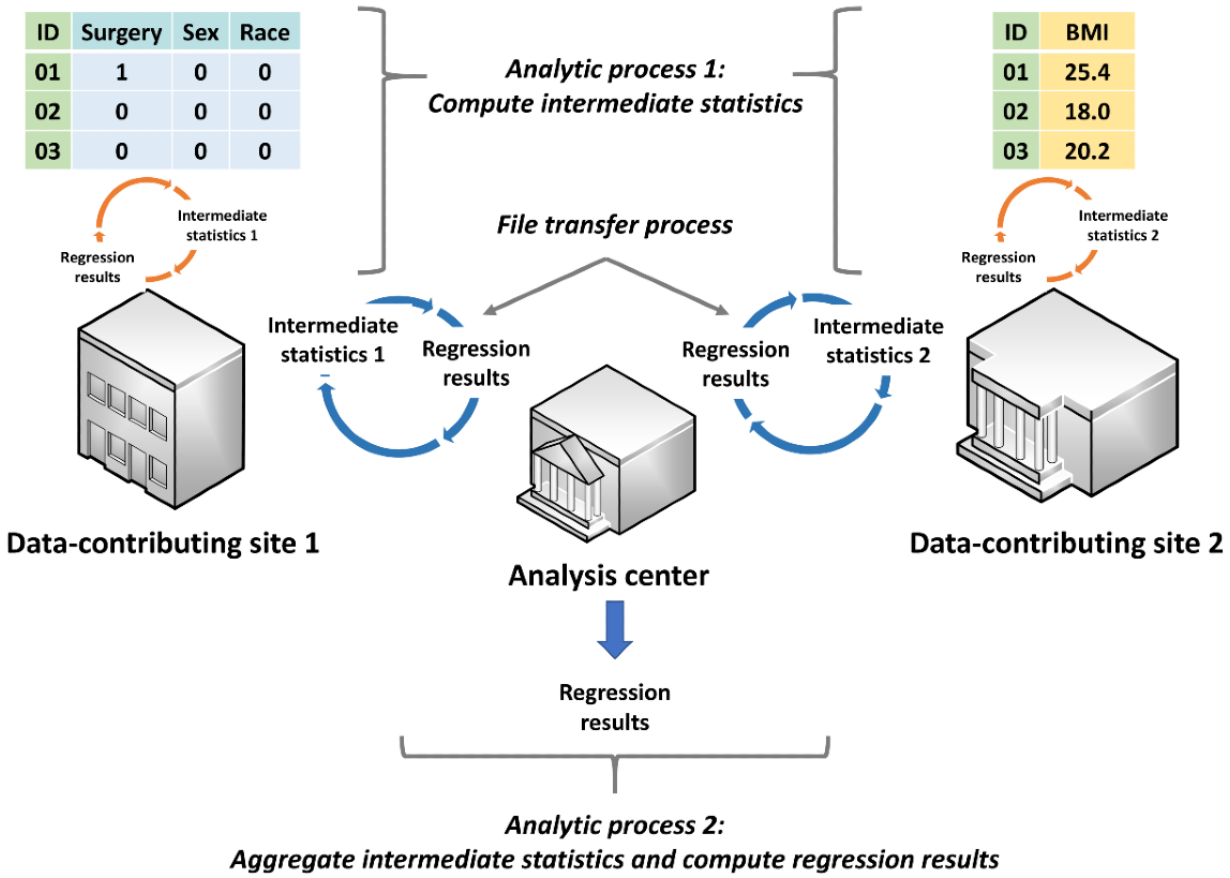
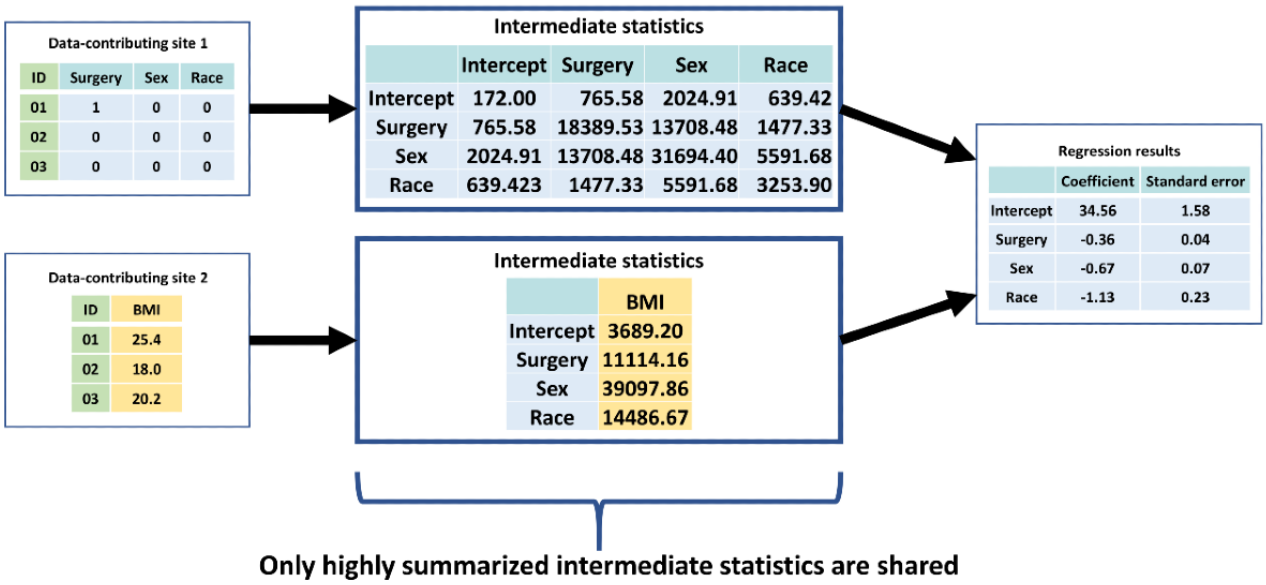


Figure 2. Distributed regression analysis in vertically partitioned data environments. In this hypothetical example, surgery, sex, and race are the independent variables, while BMI is the dependent variable. Data-contributing site 1 has data on the independent variables, while data-contributing site 2 has data on the dependent variable. Data-contributing site 2 can also have additional independent variables, as another variant of vertically partitioned data environments.



A promising privacy-protecting analytic method for both horizontally and vertically partitioned data is distributed regression, a suite of methods that perform multivariable-adjusted regression analysis with only highly summarized intermediate statistics of the patient-level data from the data-contributing sites (Figure 3) [11-13]. We have previously developed an SAS-based distributed regression analysis (DRA) package integrated with PopMedNet, an open-source file transfer software [14-17], to perform automatable distributed regression within horizontally partitioned data environments (horizontal DRA [hDRA]). We successfully tested the application in a real-world setting [18]. The feasibility of using PopMedNet to facilitate DRA with vertically partitioned data (vertical DRA [vDRA]) has not been evaluated. In this article, we describe the feasibility of using PopMedNet and enhancements to PopMedNet to facilitate automatable vDRA to protect patient privacy and support clinical research in real-world settings.

Figure 3. A hypothetical example of intermediate statistics shared by data-contributing sites in distributed linear regression analysis of vertically partitioned data. In this hypothetical example, surgery, sex, and race are the independent variables, while BMI is the dependent variable. Data-contributing site 1 has data on the independent variables, while data-contributing site 2 has data on the dependent variable.



Methods

Required Processes for Distributed Regression Analysis

Distributed regression in both horizontally and vertically partitioned environments requires two distributed analytic processes and a file transfer process (Figures 1 and 2) [11-13]. The first analytic process, which occurs at the data-contributing sites, involves computing and sharing the intermediate statistics of patient-level data with other data-contributing sites or a semitrusted third-party analysis center. The second analytic process, which occurs at the analysis center, involves aggregating the intermediate statistics and computing regression parameter estimates, standard errors, model fit statistics, and any necessary graphs (collectively called regression results hereafter). For some regression model types (eg, logistic and Cox proportional hazards), these two processes are iterative and continue until a prespecified convergence criterion is met or a prespecified maximum number of iterations is reached. When one of these prespecified conditions is fulfilled, the regression results are retained as the final results and the analysis is completed. Otherwise, the updated regression results are shared with the data-contributing sites and used to further refine the intermediate statistics. Manual transfer of the intermediate statistics can be cumbersome and error-prone. Alternatively, a semiautomated or fully automated file transfer process can be used to facilitate the iterative transfer of the intermediate statistics and regression results between the participating parties.

Implementation of hDRA Using SAS and PopMedNet

We have previously developed an SAS-based DRA package to perform the two distributed analytic processes for hDRA in a real-world setting [11,12]. This package requires an analysis center to facilitate its execution. We developed an automatable file transfer process during the execution of the DRA package by enhancing PopMedNet [1], an open-source file transfer software currently used by several large, distributed data networks, to securely transfer files through a locally installed Microsoft Windows application known as a DataMart Client using HTTPS/SSL/TLS connections [19]. PopMedNet ensures that only approved data queries are requested. Authenticated data-contributing sites are only able to transfer files to the analysis center and not to each other. All file transfers are managed by a web-based portal accessible only by the analysis center.

We integrated the SAS-based DRA package with PopMedNet to create a DRA application. The DRA application can perform distributed linear, logistic, and Cox proportional hazards regression analysis [18]. We were able to compute regression results through DRA that were precise (difference <10-12) to the regression results from the corresponding pooled patient-level data analyses using standard SAS procedures. We were also able to generate model fit graphics that were similar to the graphics obtained from the corresponding patient-level regression analysis using standard SAS procedures. With a sample size of 5452 patients, each regression model type required less than 20 minutes to complete, with the file transfer time accounting for approximately 90% of the total execution time. The hDRA application did not require participating data-contributing sites to install new software or substantially modify their hardware configuration because PopMedNet and SAS were already installed on their systems.

Feasibility of Performing vDRA Using PopMedNet

The distributed matrix computations are more complex and computationally more intensive in vDRA than in hDRA [20]. Data-contributing sites are also required to share more granular summary-level information in vDRA, leading to larger files being transferred. In hDRA, the modeling process is comprised of data-contributing sites computing the intermediate statistics and the analysis center aggregating the intermediate statistics

and computing the regression results. Most of these computations can be completed in parallel. In contrast, numerous parts of the vDRA modeling process are sequential. Specifically, the data-contributing sites must first compute components of the intermediate statistics and then compute the intermediate statistics from the components. For example, to compute the intermediate statistic known as the global covariance matrix (, where denotes a matrix of covariates) for vDRA, the data-contributing sites must first compute their site-specific covariance matrix (, where *k* denotes data-contributing site *k*) and then the off-diagonal components () of the global covariance matrix using a sequential and secure matrix multiplication algorithm (Figure 4) [20]. Once all of the components are computed, the analysis center then aggregates the intermediate statistics and computes the regression results.

Figure 4. Processes required to perform distributed linear regression in horizontally or vertically partitioned data environments with two data-contributing sites and an analysis center.

Horizontal

$$X^T X = X_1^T X_1 + X_2^T X_2,$$

where X denotes a matrix of covariates, and $X^T X$ denotes the intermediate statistics known as the global covariance matrix.

1. In a parallel process, data-contributing site 1 computes $X_1^T X_1$ and data-contributing site 2 computes $X_2^T X_2$ from their patient-level data; both then share the intermediate statistic with each other or an analysis center. $X_1^T X_1$ and $X_2^T X_2$ can be computed in parallel, because they do not require information from the other data-contributing site.
2. The analysis center computes $X^T X$ from the intermediate statistics $X_1^T X_1$ and $X_2^T X_2$.

Vertical

$$X^T X = \begin{bmatrix} X_1^T X_1 & X_1^T X_2 \\ (X_1^T X_2)^T & X_2^T X_2 \end{bmatrix},$$

where X denotes a matrix of covariates, and $X^T X$ denotes the intermediate statistics known as the global covariance matrix.

1. In a parallel process, data-contributing site 1 computes $X_1^T X_1$ and data-contributing site 2 computes $X_2^T X_2$ from their patient-level data; both then share the intermediate statistic with an analysis center. $X_1^T X_1$ and $X_2^T X_2$ can be computed in parallel, because they do not require information from the other data-contributing site.
2. In a sequential process, data-contributing sites 1 and 2 and the analysis center compute $X_1^T X_2^T$ using a secure matrix multiplication algorithm. $X_1^T X_2^T$ cannot be computed in parallel, because it requires information from the two data-contributing sites.
3. The analysis center computes $X^T X$ from the intermediate statistics submatrices $X_1^T X_1$, $X_2^T X_2$, and $X_1^T X_2^T$.

PopMedNet was not designed to optimally support these differences. Thus, we gathered the statistical and informatic requirements of vDRA and performed a feasibility analysis of using PopMedNet to facilitate vDRA with an automatable file transfer process. We used the results of this analysis to identify, develop, and implement enhancements to PopMedNet to improve its technical capability to facilitate automatable vDRA. As guiding principles, we required enhancements to support execution times that lasted for only a few hours, to not disrupt existing PopMedNet workflows, to be developed and implemented with minimal to moderate effort at the data-contributing sites, and to not lead to major modifications to hardware configurations or new software installations at the data-contributing sites.

Results

Findings From Assessment of the Existing PopMedNet Functionalities to Facilitate Automatable vDRA

With the prior enhancements to PopMedNet (version 6.7) to facilitate automatable hDRA, we had the necessary technical infrastructure to perform automatable vDRA. However, vDRA would be limited to analysis of small cohort sizes and regression models with few covariates. As described above, vDRA requires matrix computations that are more complex and computationally more intensive than hDRA, data-contributing sites to share more granular information and files of larger sizes, and a modeling process that is mostly sequential. These differences would increase computation time and file transfer time, which would lead to considerably longer execution times (Table 1). Moreover, executing vDRA with a large cohort of patients or a large number of covariates would further increase execution time and render the PopMedNet configuration developed for hDRA impractical. Additional enhancements to PopMedNet were necessary to ensure that DRA is a feasible analytic option for vertically partitioned data.

Table 1. Differences between horizontal and vertical distributed regression analysis and their impacts on execution time.

Component of analysis	Distributed regression analysis		Impact on execution time in vertical distributed regression analysis[a]
	Horizontal	Vertical	
Computation sequence[b]	1. Computes intermediate statistics 2. Computes regression results[c]	1. Computes components of the intermediate statistics 2. Computes intermediate statistics 3. Computes regression results[c]	Increases computation and file transfer times
Computation process	Most computations can be completed in a parallel process	Most computations require a sequential process	Increases computation and file transfer times
File transfer sizes	Kilobytes to megabytes [18]	Kilobytes to infinity[d]	Increases file transfer times
Dimension of an example matrix transfer[e]	$p \times p$	$n \times n$	Increases file transfer times

[a]Execution time in horizontal distributed regression analysis serves as the baseline.

[b]Sequence of computations required to compute regression results; the sequence is iterative for some regression model types.

[c]Include regression parameter estimates, standard errors, model fit statistics, and model fit graphics.

[d]File sizes increase as cohort size or number of covariates in the regression model increases.

[e]p is the number of regression model covariates; n is the number of observations at each data-contributing site.

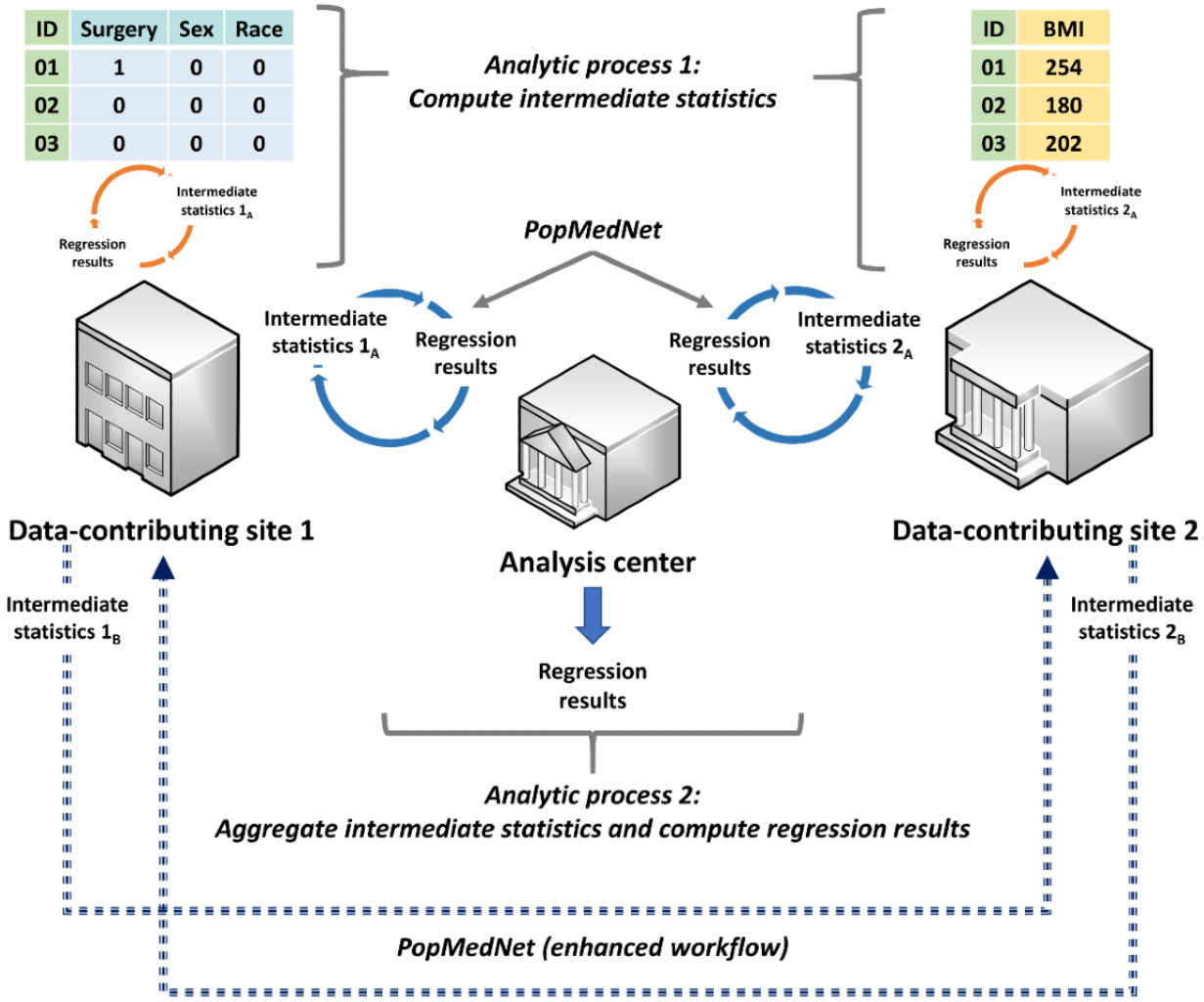
Implemented Enhancements to PopMedNet to Enable Automatable and Efficient Implementation of vDRA

We identified and implemented two enhancements to PopMedNet to improve its technical capability to facilitate automatable vDRA in the real-world setting. These enhancements decrease execution time, require only minimal or moderate development and implementation efforts, do not disrupt existing PopMedNet workflows, and do not require data-contributing sites to modify their hardware configurations.

First, we enhanced PopMedNet to allow simultaneous upload and download of multiple files. The previous version of PopMedNet only allowed one file to be uploaded and downloaded at a given time. Concurrent transfer of multiple files decreases file transfer time. Second, we enhanced PopMedNet to allow direct file transfers between data-contributing sites (Figure 5). Previously, PopMedNet did not allow direct file transfers between data-contributing sites. File transfer was only possible between the analysis center and the data-contributing sites; any files to be shared between data-contributing sites first had to be transferred through the analysis center. This design aimed to reduce the potential risk of two or more data-contributing sites colluding against the other sites to derive information about specific patients using the summary statistics, but it doubled the file transfer times. Allowing direct file transfers between the data-contributing sites would enable vDRA algorithms that are simpler, share less granular information, share smaller files, and require fewer file transfers. It would also allow more computations to be done in parallel, which would decrease computation and file transfer times.

To minimize the risk of collusion under the enhanced file transfer process, we implemented a trust matrix, where the analysis center can prespecify and govern the file transfer process between data-contributing sites. Only the analysis center will have access and permission to modify the trust matrix. Any data-contributing sites that violate the trust model will prompt PopMedNet to terminate the analysis.

Figure 5. Enhanced PopMedNet workflow to enable automatable distributed regression analysis in vertically partitioned data environments. In this hypothetical example, surgery, sex, and race are the independent variables, while BMI is the dependent variable. Data-contributing site 1 has data on the independent variables, while data-contributing site 2 has data on the dependent variable.



Discussion

Principal Findings

PopMedNet can facilitate the execution of automatable vDRA. We identified and implemented two enhancements to PopMedNet that improved its technical capability to facilitate vDRA in real-world settings. The first enhancement was concurrent uploading and downloading of multiple files and the second involved enhancing the PopMedNet trust model to allow certain prespecified and preapproved file transfer processes between the data-contributing sites. Both enhancements decrease the file transfer and computation times needed to perform vDRA, which limit connectivity issues (eg, firewall timeouts, network connections, and connections to the internet) as a barrier to performing vDRA in real-world settings. Connectivity may seem inconsequential with high-speed internet and high-performing computers, but issues of connectivity are compounded when there are multiple parties participating in a regression analysis that requires numerous iterations and file transfers.

Other Research in vDRA

There have been efforts to make vDRA a practical analytic option in real-world settings. Li and colleagues [3] developed VERTIcal Grid lOgistic regression (VERTIGO) to perform distributed logistic regression analysis in vertically partitioned data. Dai and colleagues [4] recently developed VERTICOX to perform distributed Cox proportional hazards regression analysis in vertically partitioned data. Both studies found that cohort size influenced the operational performance of their methods. Similar to the findings in our feasibility analysis, Li and colleagues [3] concluded that vDRA performed with VERTIGO was limited to analyzing relatively small cohort sizes. These authors identified several potential solutions to improve the operational performance of VERTIGO, including performing the matrix computations on graphic processing units, parallelizing the matrix computations on multiple cores or machines, and dividing the matrix computations into smaller parts (a divide-and-conquer strategy). Performing vDRA on graphic processing units may require new hardware, which would violate our guiding principles. We explored parallelizing the matrix computations in our feasibility analysis but concluded that it would require meaningful changes to the hardware

configuration and installation of new software at the data-contributing sites. We explored implementing a similar divide-and-conquer strategy in the design of our vDRA algorithms by first horizontally partitioning the vertically partitioned data sets into smaller blocks and then performing the required matrix computations within the blocks. This design should decrease computation times because the computations are completed with smaller matrices.

Limitations

Our two enhancements improved PopMedNet's technical capability to facilitate vDRA. However, there are several factors that may limit the use of PopMedNet to enable automatable vDRA in real-world settings. First, real-world implementation of PopMedNet-supported vDRA will be driven by the degree of automation between the participating data-contributing sites and the file transfer speed of the slowest site. Numerous parts of the vDRA modeling process are executed sequentially. Thus, if a site performs vDRA with a manual file transfer process or has a slow file transfer speed, the overall execution time will be limited by the response time of the slowest site. This may deter data-contributing sites from using vDRA or PopMedNet to analyze data from multiple data sources.

Second, we could only implement enhancements to PopMedNet that adhered to our guiding principles, which were developed based on existing PopMedNet users. There may be unforeseen challenges that require major enhancements or changes to the PopMedNet topology and infrastructure if vDRA were to be conducted in other data-contributing sites with different software and hardware configurations.

Third, the acceptance of automatable vDRA and enhanced PopMedNet capabilities by data-contributing sites should not be overlooked. In our previous experience, data-contributing sites were reluctant to use the fully automated PopMedNet workflow to facilitate hDRA. It was only after an initial roll-out phase with the manual and semiautomated workflows and confirmation that the intermediate statistics did not contain identifiable patient information that they were willing to experiment with the fully automated PopMedNet workflow. With vDRA requiring the transfer of more granular information and longer computation and file transfer times than hDRA, some data-contributing sites may require a similar roll-out phase to build trust and acceptance of automatable vDRA.

Fourth, we chose to build upon our previous work and enhanced PopMedNet to facilitate vDRA. This may limit our vDRA package to organizations who use PopMedNet as their file transfer software. However, PopMedNet is currently used by several large distributed data networks, including the Sentinel System [14], the National Patient-Centered Clinical Research Network [15], and the National Institutes of Health Health Care Systems Research Collaboratory [16]. These networks have established infrastructure (eg, harmonized data, data use agreements, governance) and processes that can streamline the use of vDRA to analyze data from multiple data sources. Thus, these networks can readily implement PopMedNet-supported hDRA and vDRA.

Future Work

With the enhancements to PopMedNet, we have started implementing the divide-and-conquer vDRA algorithms that leverage PopMedNet's new functionality to transfer files between data-contributing sites. We are also exploring enhancements to the vDRA algorithms to reduce the number of files needed to be transferred. Reducing the number of files transferred will decrease the overall file transfer time and make vDRA a practical analytic option in real-world settings. To offer better privacy protection, we are only implementing vDRA algorithms that limit the potential for back calculations. We will explore the feasibility of combining our vDRA algorithms with data manipulation techniques, such as perturbation and encryption, to provide additional layers of protection [5]. We plan to integrate these vDRA algorithms with PopMedNet to create a vDRA application and test it with real-world data.

Conclusion

PopMedNet can be used to facilitate automatable vDRA. We have implemented two enhancements to the PopMedNet workflow to improve its technical capability to facilitate vDRA in real-world settings. PopMedNet has the potential to increase clinical research and collaboration across multiple data-contributing sites while protecting patient privacy and proprietary interests of the data-contributing sites.

Acknowledgments

This work was supported by the National Institute of Biomedical Imaging and Bioengineering (U01EB023683) and the National Center for Advancing Translational Sciences (UL1 TR002014) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

None declared.

References

1. Her QL, Malenfant JM, Malek S, Vilk Y, Young J, Li L, et al. A Query Workflow Design to Perform Automatable Distributed Regression Analysis in Large Distributed Data Networks. EGEMS (Wash DC) 2018 May 25;6(1):11 [FREE Full text] [doi: 10.5334/egems.209] [Medline: 30094283]

2. Wolfson M, Wallace SE, Masca N, Rowe G, Sheehan NA, Ferretti V, et al. DataSHIELD: resolving a conflict in contemporary bioscience--performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010 Oct;39(5):1372-1382 [FREE Full text] [doi: 10.1093/ije/dyq111] [Medline: 20630989]
3. Li Y, Jiang X, Wang S, Xiong H, Ohno-Machado L. VERTIcal Grid lOgistic regression (VERTIGO). J Am Med Inform Assoc 2016 May;23(3):570-579 [FREE Full text] [doi: 10.1093/jamia/ocv146] [Medline: 26554428]
4. Dai W, Jiang X, Bonomi L, Li Y, Xiong H, Ohno-Machado L. VERTICOX: Vertically Distributed Cox Proportional Hazards Model Using the Alternating Direction Method of Multipliers. IEEE Trans Knowl Data Eng 2020 Apr 01:1-1. [doi: 10.1109/tkde.2020.2989301]
5. El Emam K, Samet S, Arbuckle L, Tamblyn R, Earle C, Kantarcioglu M. A secure distributed logistic regression protocol for the detection of rare adverse drug events. J Am Med Inform Assoc 2013 May 01;20(3):453-461 [FREE Full text] [doi: 10.1136/amiajnl-2011-000735] [Medline: 22871397]
6. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. Med Care 2013 Aug;51(8 Suppl 3):S4-10. [doi: 10.1097/MLR.0b013e31829b1bb1] [Medline: 23752258]
7. Toh S, Reichman ME, Houstoun M, Ding X, Fireman BH, Gravel E, et al. Multivariable confounding adjustment in distributed data networks without sharing of patient-level data. Pharmacoepidemiol Drug Saf 2013 Nov;22(11):1171-1177. [doi: 10.1002/pds.3483] [Medline: 23878013]
8. Toh S, Shetterly S, Powers JD, Arterburn D. Privacy-preserving analytic methods for multisite comparative effectiveness and patient-centered outcomes research. Med Care 2014 Jul;52(7):664-668. [doi: 10.1097/MLR.0000000000000147] [Medline: 24926715]
9. Slavkovic A, Nardi Y, Tibbits M, editors. Secure logistic regression of horizontally vertically partitioned distributed databases. 2007 Presented at: Seventh IEEE International Conference on Data Mining Workshops; March 28, 2007; Omaha, NE, USA. [doi: 10.1109/icdmw.2007.114]
10. Karr A, Lin X, Sanil A, Reiter J. Privacy-preserving analysis of vertically partitioned data using secure matrix products. J Off Stat 2009;25(1):125.
11. Her Q, Vilk Y, Young J, Zhang Z, Malenfant J, Malek S. A distributed regression analysis application based on SAS software. Part I: Linear and logistic regression. arxiv.org. 2018 Aug 07. URL: https://arxiv.org/abs/1808.02387 [accessed 2021-03-30]
12. Vilk Y, Zhang Z, Young J, Her Q, Malenfant J, Malek S. A distributed regression analysis application based on SAS software. Part II: Cox proportional hazards regression. arxiv.org. 2018 Aug 07. URL: https://arxiv.org/abs/1808.02392 [accessed 2021-03-30]
13. Karr AF, Feng J, Lin X, Sanil AP, Young SS, Reiter JP. Secure analysis of distributed chemical databases without data integration. J Comput Aided Mol Des 2005;19(9-10):739-747. [doi: 10.1007/s10822-005-9011-5] [Medline: 16267693]
14. Platt R, Carnahan RM, Brown JS, Chrischilles E, Curtis LH, Hennessy S, et al. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. Pharmacoepidemiol Drug Saf 2012 Jan;21 Suppl 1:1-8. [doi: 10.1002/pds.2343] [Medline: 22262586]
15. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;21(4):578-582 [FREE Full text] [doi: 10.1136/amiajnl-2014-002747] [Medline: 24821743]
16. Richesson RL, Hammond WE, Nahm M, Wixted D, Simon GE, Robinson JG, et al. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. J Am Med Inform Assoc 2013 Dec;20(e2):e226-e231 [FREE Full text] [doi: 10.1136/amiajnl-2013-001926] [Medline: 23956018]
17. PopMedNet. URL: https://www.popmednet.org/ [accessed 2021-03-31]
18. Her Q, Malenfant J, Zhang Z, Vilk Y, Young J, Tabano D, et al. Distributed Regression Analysis Application in Large Distributed Data Networks: Analysis of Precision and Operational Performance. JMIR Med Inform 2020 Jun 04;8(6):e15073 [FREE Full text] [doi: 10.2196/15073] [Medline: 32496200]
19. Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood) 2014 Jul;33(7):1178-1186. [doi: 10.1377/hlthaff.2014.0121] [Medline: 25006144]
20. Karr AF, Fulp WJ, Vera F, Young SS, Lin X, Reiter JP. Secure, Privacy-Preserving Analysis of Distributed Databases. Technometrics 2007 Aug 01;49(3):335-345. [doi: 10.1198/004017007000000209]

Abbreviations

DRA: distributed regression analysis
hDRA: horizontal distributed regression analysis
vDRA: vertical distributed regression analysis
VERTIGO: VERTIcal Grid lOgistic regression

Edited by C Lovis; submitted 18.06.20; peer-reviewed by D Tabano, Z Zhang; comments to author 21.08.20; revised version received 25.01.21; accepted 07.03.21; published 23.04.21.

Please cite as:
Her Q, Kent T, Samizo Y, Slavkovic A, Vilk Y, Toh S
Automatable Distributed Regression Analysis of Vertically Partitioned Data Facilitated by PopMedNet: Feasibility and Enhancement Study
JMIR Med Inform 2021;9(4):e21459
URL: https://medinform.jmir.org/2021/4/e21459
doi:10.2196/21459
PMID:33890866

Original Paper

Mortality Prediction of Patients With Cardiovascular Disease Using Medical Claims Data Under Artificial Intelligence Architectures: Validation Study

Linh Tran[1], MSc; Lianhua Chi[2], PhD; Alessio Bonti[1], PhD; Mohamed Abdelrazek[1], PhD; Yi-Ping Phoebe Chen[2], PhD

[1]School of Info Technology, Deakin University, Burwood, Australia

[2]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Australia

Corresponding Author:
Lianhua Chi, PhD
Department of Computer Science and Information Technology
La Trobe University
Beth Gleeson Bldg, 2rd Fl, #242
La Trobe University
Bundoora, 3086
Australia
Phone: 61 94792454
Email: l.chi@latrobe.edu.au

Abstract

Background: Cardiovascular disease (CVD) is the greatest health problem in Australia, which kills more people than any other disease and incurs enormous costs for the health care system. In this study, we present a benchmark comparison of various artificial intelligence (AI) architectures for predicting the mortality rate of patients with CVD using structured medical claims data. Compared with other research in the clinical literature, our models are more efficient because we use a smaller number of features, and this study could help health professionals accurately choose AI models to predict mortality among patients with CVD using only claims data before a clinic visit.

Objective: This study aims to support health clinicians in accurately predicting mortality among patients with CVD using only claims data before a clinic visit.

Methods: The data set was obtained from the Medicare Benefits Scheme and Pharmaceutical Benefits Scheme service information in the period between 2004 and 2014, released by the Department of Health Australia in 2016. It included 346,201 records, corresponding to 346,201 patients. A total of five AI algorithms, including four classical machine learning algorithms (logistic regression [LR], random forest [RF], extra trees [ET], and gradient boosting trees [GBT]) and a deep learning algorithm, which is a densely connected neural network (DNN), were developed and compared in this study. In addition, because of the minority of *deceased* patients in the data set, a separate experiment using the Synthetic Minority Oversampling Technique (SMOTE) was conducted to enrich the data.

Results: Regarding model performance, in terms of discrimination, GBT and RF were the models with the highest area under the receiver operating characteristic curve (97.8% and 97.7%, respectively), followed by ET (96.8%) and LR (96.4%), whereas DNN was the least discriminative (95.3%). In terms of reliability, LR predictions were the least calibrated compared with the other four algorithms. In this study, despite increasing the training time, SMOTE was proven to further improve the model performance of LR, whereas other algorithms, especially GBT and DNN, worked well with class imbalanced data.

Conclusions: Compared with other research in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient because we use a smaller number of features but still achieve high performance. This study could help health professionals accurately choose AI models to predict mortality among patients with CVD using only claims data before a clinic visit.

(JMIR Med Inform 2021;9(4):e25000) doi:10.2196/25000

KEYWORDS

mortality; cardiovascular; medical claims data; imbalanced data; machine learning; deep learning

Introduction

Background

In Australia, cardiovascular disease (CVD) is the most concerning health problem, killing more people than any other disease and placing heavy burdens on the health care system because of enormous costs and on individuals and the community owing to resulting disabilities. CVD was the leading cause of death among Australians in 1997, accounting for 52,641 deaths, 41% of all deaths [1]. An estimated 1.2 million (5.6%) Australian adults aged 18 years and more had one or more conditions related to heart or vascular disease, including stroke, in 2017-2018, based on self-reported data from the Australian Bureau of Statistics 2017-2018 National Health Survey. The prevalence of CVD by age group and sex, in 2017-2018, is shown in Table 1.

Table 1. Prevalence of cardiovascular disease by age group and sex, 2017-2018.

Age group (years)	Men, n[a]	Women, n[a]	Total, n[a]	Men, % (95% CI)[b]	Women, % (95% CI)[b]	Total, % (95% CI)[b]
18-44	31,400	56,600	88,000	0.7 (0.3-1.1)	1.2 (0.7-1.8)	1.0 (0.7-1.3)
45-54	50,600	42,300	92,900	3.3 (2.4-4.2)	2.6 (1.7-3.5)	3.0 (2.4-3.6)
55-64	136,700	114,700	251,500	10.0 (7.6-12.4)	7.9 (6.0-9.9)	8.9 (7.4-10.5)
65-74	208,900	135,600	344,500	19.8 (17.2-22.4)	12.2 (10.0-14.4)	15.9 (14.3-17.5)
75+	213,200	160,100	373,300	32.1 (27.1-37.0)	20.3 (17.5-23.1)	25.7 (23.1-28.2)
Persons (number/age-standardized rate[c])	640,800	509,300	1,150,200	6.5 (5.9-7.0)	4.8 (4.3-5.3)	5.6 (5.2-5.9)

[a]Due to rounding, discrepancies may occur between sums of the component items and totals.

[b]CI is a statistical term describing a range (interval) of values within which we can be "confident" that the true value lies, usually because it has a 95% or higher chance of doing so.

[c]Age-standardized to the 2001 Australian Standard Population (Source: AIHW analysis of ABS 2019).

The major risk factors for CVD are tobacco smoking, high blood pressure, high blood cholesterol, overweight, insufficient physical activity, high alcohol use, and type 2 diabetes [1]. CVD treatments are usually prescribed in combination with other drugs such as antidiabetics, antihypertensives, lipid-lowering drugs, anticoagulants, and antiplatelet agents [2]. Medication use is an important management factor for patients diagnosed with heart disease besides eating a healthy diet and maintaining fitness with regular physical activity. Medications are used to minimize symptoms, reduce the risk of exacerbation, and improve the quality of life.

Many methods have been developed to predict the mortality rate of patients with CVD by using many algorithms and predictor variables. There are 3 main methods for forecasting mortality: explanation, expectation, and extrapolation [3]. Of these, the most common basis of forecasting mortality is extrapolation, which assumes that the future state is highly correlated to the past. In the clinical literature, historical electronic health records (EHRs) are widely used to develop artificial intelligence (AI) models that can predict the health outcomes of patients. Information commonly extracted from EHR as input for AI models includes patient demographics, health indices, medical conditions, biomedical images, or clinical notes, whereas structured medical claims data are rarely used. Although medical claims data inadequately inform patient health conditions, this source of information is crucial in reflecting patient health care access frequency and level of participation in disease prevention or treatment, which has a great impact on patient health outcomes.

In this study, we present a benchmark comparison of the performance of different AI architectures: 4 classical machine learning (ML) algorithms (logistic regression [LR], random forest [RF], extra trees [ET], and gradient boosting trees [GBT]) and a deep learning algorithm, which is a densely connected neural network (DNN) that uses medical scheduling and pharmaceutical dispensing information from historical claims data to predict the mortality rate of patients with CVD. Compared with other research in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient because we use a smaller number of features but still achieve high performance. Furthermore, we also propose Synthetic Minority Oversampling Technique (SMOTE), a technique to enrich training data and handle class imbalance, as a tool to improve the performance of the developed AI models.

Related Work

Recent trends involve using AI models to learn patterns from large data sets to predict mortality with higher accuracy [4]. The American College of Cardiology Foundation's National Cardiovascular Data Entry conducted a study that used statistical analysis to predict the rate of risk in percutaneous coronary intervention. The study results show that ML models perform better in terms of accuracy than classical statistical models [5]. One study showed that ML models such as RF, decision tree, and LR perform exceptionally well owing to today's computational power, which allows them to process data from the electrical health records [6] of patients. ML models deployed on routine clinical data performed better than standard

cardiovascular risk assessment models and had great merits in terms of preventive treatment and avoidance of mistreatment for CVD according to a study conducted on a large sample of patients in the United Kingdom [7]. Moreover, using neural networks for predictive analysis of illnesses was shown to be fruitful as early as in 2005 [8]. Wang et al [9] predicted the mortality rate because of heart failure by deploying a convolutional, layered neural network that inculcated feature rearrangement to select the best features. Another study has shown that deep neural networks perform better than traditional ML models with respect to accuracy and available sample size [10].

Many factors have been considered to predict the health outcomes of patients with heart disease. Some techniques used to extract learning features include automated imaging interpretation [11,12], natural language processing or text mining [13,14], and EHR extraction [15-18]. Imaging interpretation has been carried out by using deep neural networks [12] with promising results. Natural language processing of clinical notes has been shown to be able to correctly identify risks of CVD patients [13], whereas systematic application of text mining to the EHR has had variable success in the detection of cardiovascular phenotype [14]. It has been proven that applying ML helps identify clinically relevant patterns in the data [19]. Feature extraction from EHR allows the use of many factors, such as patient demographics, characteristics, and health conditions, including cardiovascular health (CVH) indices [20] or percutaneous coronary intervention indices [16,17] in predicting mortality risks.

On the basis of these studies, the mortality rate of patients in the cardiology cohort has been accurately predicted using a variety of algorithms, methods, and predictor features. However, there has been little focus on using medical claims to predict the health outcomes of patients with CVD. This information reflects patient medication usage, health care access frequency, and level of participation in disease prevention or treatment, which have a great impact on the determination of patient health outcomes [21]. Hence, to close this literature gap, in this study, mortality will be predicted based on patient medical schedule information and pharmaceutical dispensing history acquired from medical claims.

The Pharmaceutical Benefits Scheme (PBS) and Medicare Benefits Schedule (MBS) claims data collected by the Department of Human Services and held by the Department of Health have great potential to provide further insight into the medical scheduling and pharmaceutical dispensing history of patients with CVD. This study uses the PBS and MBS claims data in the period between 2004 and 2014 to investigate the mortality rate of patients with heart disease conditions in Australia and to build and compare 5 AI models to predict the mortality risk of a patient under these conditions. We built prediction models based on the patient's age, gender, relevant medication prescriptions, medical schedule information, and pharmaceutical dispensing history obtained from the data set. We then assessed and compared the performance of each model and suggested recommendations for future work.

Objectives

The primary aim of this research is to support health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit. Compared with other research in the clinical literature involving AI models using claims data to predict patient health outcomes, our models are more efficient because we use a smaller number of features but still achieve high performance. This study has applications in supporting health clinicians to accurately predict mortality among patients with CVD using only claims data before a clinic visit.

Methods

AI Architectures

In this study, 4 classical ML algorithm architectures, LR, RF, ET, and GBT, along with a deep learning algorithm called DNN were used to develop mortality prediction models. The MBS and PBS data sets are well structured and very informative and allows simple algorithms to learn better. Because our study deals with a probabilistic prediction problem, we put more emphasis on the discrimination and calibration of the model performance. Through initial experiments we found that LR, RF, ET, and GBT are classical ML algorithms that produce the best performance in terms of these two criteria. On the other hand, we were curious about how a state-of-the-art deep learning algorithm might perform on the data set. We developed the simplest neural network, a DNN, for further comparison and insights. We chose not to develop more complex deep learning architectures such as RNN or CNN because these algorithms are not necessary for such structured data sets to perform well. In this section, these experimental algorithms are described and their architectures proposed.

Logistic Regression

LR is a supervised ML algorithm. It is a powerful and well-established method for binary classification problems [22]. LR is extended based on linear regression and can be used to calculate the probability of an event that has 2 possible outcomes by assigning weights to a number of predictor variables (features). Given a set of independent variables

$$x_1,x_2,x_3,\ldots,x_n \quad \textbf{(1)}$$

and a dependent variable *y*, which takes values between 0 and 1, first, LR is designed to find a set of weights

$$b_1,b_2,b_3,\ldots,b_n \quad \textbf{(2)}$$

for each of the independent variables so that the following linear equation outputs a logit score:

$$\text{logit} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_nx_n \quad \textbf{(3)}$$

From this logit score, probability y is then derived by the following formula:

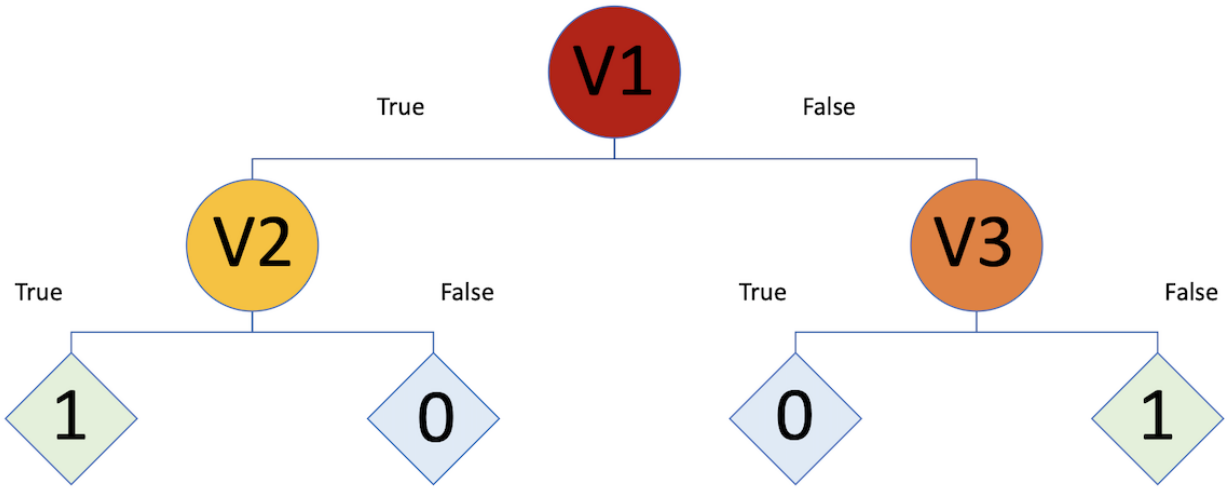
To use the LR as a binary classifier, a threshold must be assigned to differentiate between 2 classes. Normally, LR will classify an input instance with P>.50 as a positive class; otherwise, it is

classified as a negative class. Depending on the problem, 0 and 1 can be translated into different meanings.

Random Forest

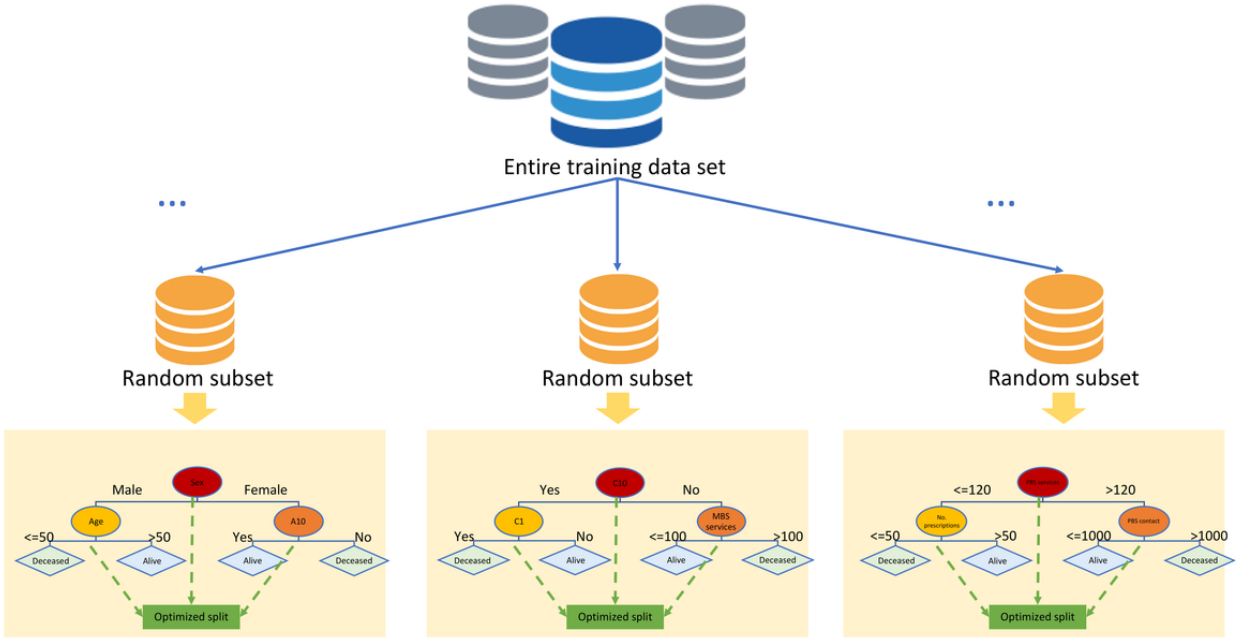
Before describing the RF algorithm, it is important to understand the concept of the decision tree algorithm [23]. DT is one of the simplest and earliest ML algorithms. It structures the decision logic into a tree-like model. The nodes in a DT tree are partitioned into different levels, where the uppermost node is called the root node, whereas other nodes that have at least one child represent tests on input variables/features [24]. Depending on some criterion of the test, higher nodes are split into lower nodes repeatedly toward the leaf nodes [25], which have no child at all and correspond to the decision outcomes. An illustration of a simple DT is shown in Figure 1. According to Figure 1, the 3 circles -Sex, Age, and *A10*- are tested on the corresponding input variables, whereas the rhombuses at the end are the classification outcomes (*deceased* or *alive*).

Figure 1. Decision tree.



An RF is an ensemble classifier consisting of many DTs similar to a forest with many trees [26]. Different DTs in an RF are trained using different parts of the training data set and tested on different subsets of input variables. To classify a new instance, the input vector of the instance is pushed through each DT in the forest. Each DT makes decisions on a different part of the input vector and provides a classification outcome. The forest then makes a final prediction by majority vote in classification problems and by arithmetic average in regression problems. Because the RF algorithm aggregates outcomes from many different DTs to make a decision, the result has a smaller variance compared with the consideration of a single DT for the same data set. In addition, similar to other tree-based ensembles, variables for each tree in RF are randomized, whereas node-splitting cut points are locally optimized according to the criterion [26]. Figure 2 illustrates the RF algorithm. As shown in Figure 2, the training data set is randomly split into the desired number of trees in the forest, and each random subsample is then used to train a decision tree that is tested on a randomly selected subset of input variables.

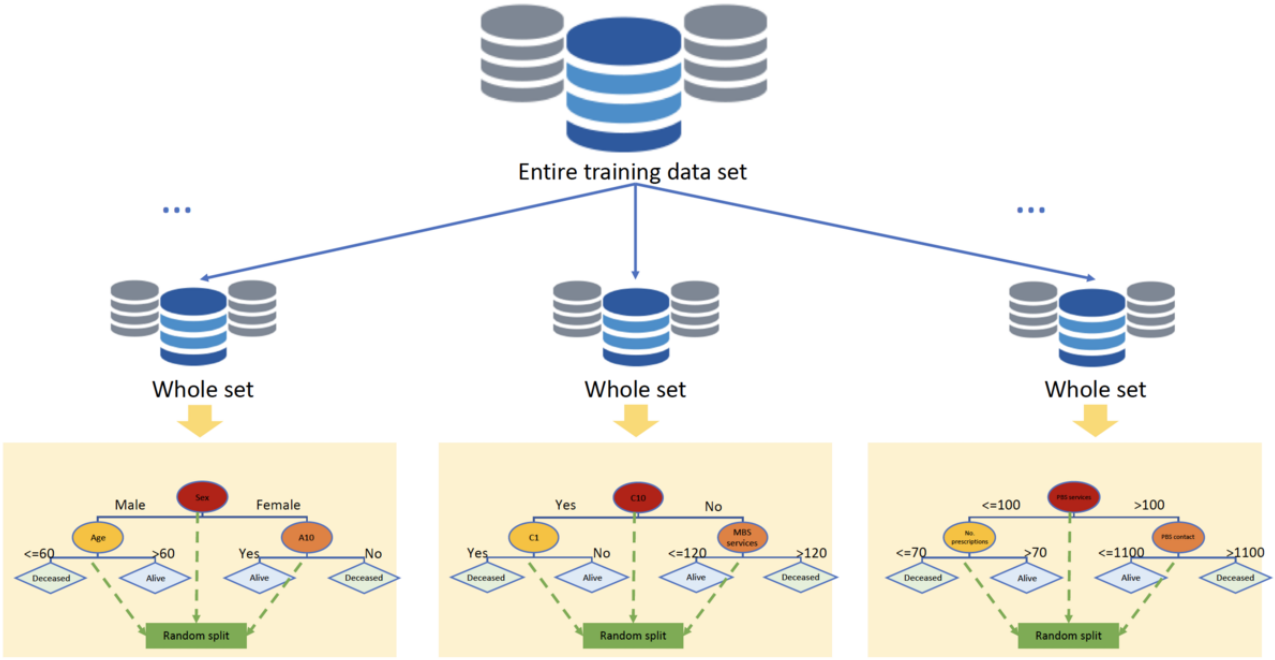
Figure 2. Random forests.



Extra Trees

The extremely randomized trees or the ET algorithm is also an ensemble classifier consisting of many single DTs similar to RF. The ET method also uses a random subset of features to train each base estimator [27]. However, the two main differences between RF and other tree-based ensemble methods are that RF splits nodes by choosing cut points fully at random (or random selection of threshold), and RF uses the whole learning sample to grow each tree in the ensemble rather than a subset of training data [28]. The final prediction produced is the aggregate of the predictions of all trained trees, yielded by the majority vote or arithmetic average in classification problems or regression problems, respectively. In terms of bias variance, ET is able to reduce the variance more effectively than the weaker randomization schemes used by other ensemble methods. On the other hand, a full training sample rather than bootstrap batches is used to train each base estimator in an attempt to minimize bias [28]. A simple illustration of the ET model is shown in Figure 3.

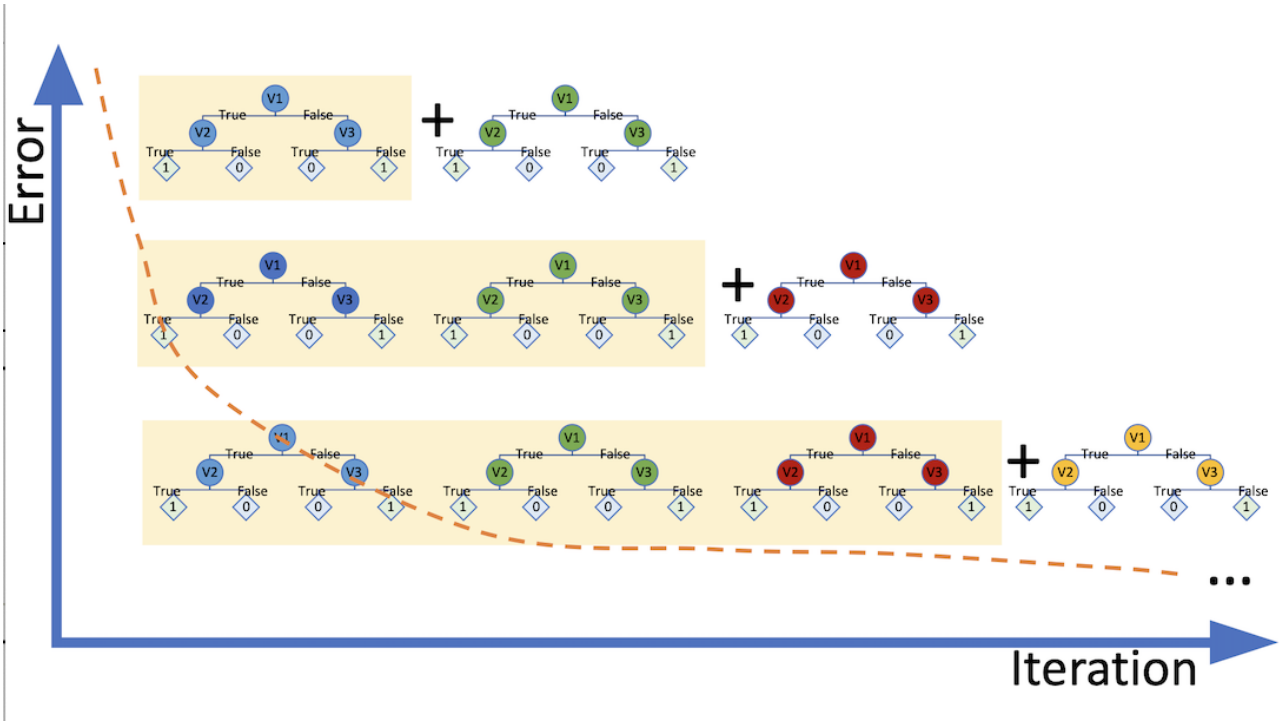
Figure 3. Extra trees.



Gradient Boosting Trees

GBT is another popular ML algorithm that uses a tree-based ensemble method, and was first proposed by Friedman [29]. This approach trains learners (decision trees) by minimizing the loss function, which is computed using the gradient descent method [30]. To train a GBT, the algorithm first selects a very simple decision tree from the learning sample with equal weights. On the basis of the results of this weak learner, it tries to create a new learner who assigns higher weights to nodes that are more difficult to split and lower weights to those that are easier to split [30]. By doing this, the new learner is able to minimize the errors of the previous learner. As this process continues, the loss function is optimized [29], making each new model have a better goodness of fit with the observation data. Figure 4 illustrates the mechanism of the GBT algorithm.

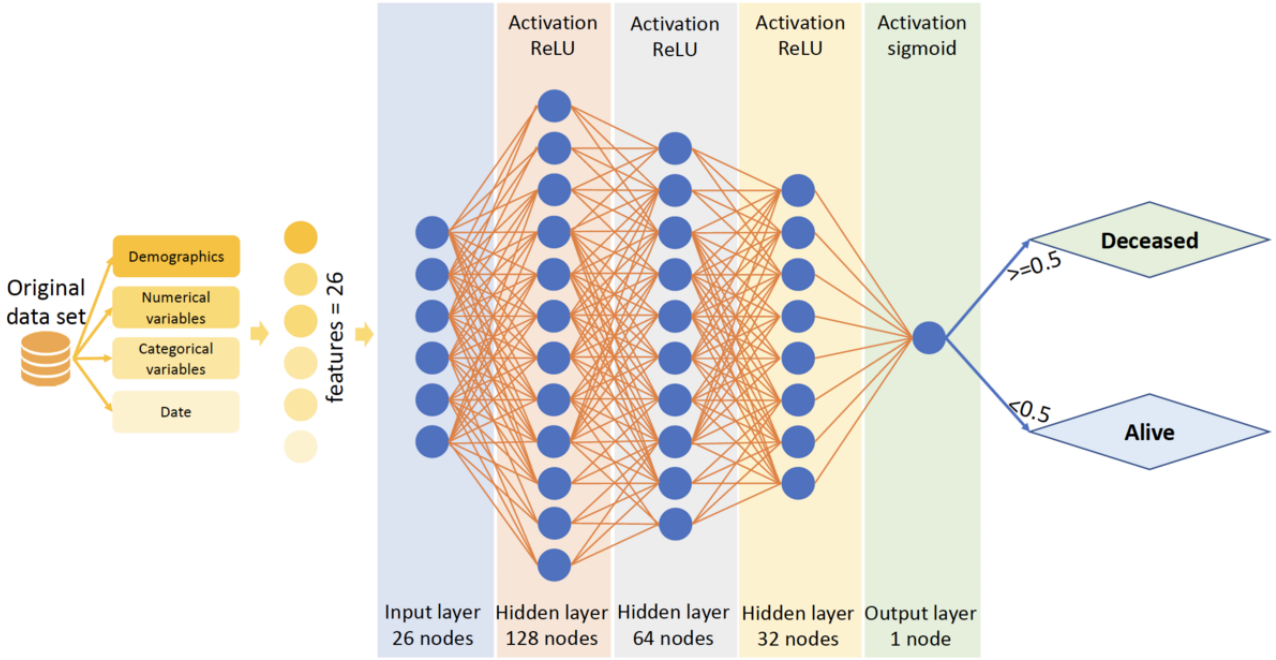
Figure 4. Gradient boosting trees workflow.



Densely Connected Neural Network

An artificial neural network (ANN) [3] is a deep learning architecture that replicates the neuron system inside the human brain. McCulloch and Pitts [31] first proposed ANN, and the concept was later popularized by the research work of Rumelhart et al [32]. In the human brain, neurons are linked together by numerous axon connections [33] and are responsible for adapting, processing, and storing information toward (inputs) and away (outputs) from the brain. Likewise, an ANN has hundreds or even thousands of artificial neurons called processing units, which are interconnected by nodes. In the ANN architecture, nodes are grouped into layers, depending on the activation they implement on the data. In the ANN, the output of one node is the input to another node. Subsequently, the input node after receiving information from the previous output node, based on an internal weighting system, attempts to produce the next output. Through repeated training, the weight system can amplify or weaken the level of communication between nodes. After mature training, which optimizes the weight system, a trained ANN can predict the test data. Because ANNs can be constructed by many layers and neurons, this method is considered a deep learning algorithm. Many types of ANNs are currently used in the literature, including feedforward neural networks, recurrent neural networks, convolutional neural networks, and modular neural networks. In this study, because our input data are well structured, allowing a neural network to learn effectively, we present the simplest form of ANN, which is a DNN. Figure 5 shows an illustration of the proposed DNN with 3 hidden layers.

Figure 5. Artificial neural network architecture. ReLU: Rectified Linear Unit.



Results

Benchmark Data

On August 1, 2016, the Department of Health released approximately 1 billion lines of anonymous historical health data relating to approximately 3 million Australians on data.gov.au. The information released includes details on medical services provided to Australians by health professionals, along with details of subsidized information. Claims data for a random 10% sample of Australians are made available for research institutions, health professionals, and universities. The data release includes historical medicare data (from 1984) and PBS data (from 2003) up to 2014. The release comprises 2 files corresponding to the 2 types of service information (MBS and PBS) and a separate patient demographic file. The data set used in this study was obtained from the MBS and PBS service information and patient demographic data by patient IDs. It originally included 346,201 records corresponding to 346,201 patients; however, 19 patients who had inadequate information were removed. Following this exclusion, the final data set comprised a total of 346,182 patients.

The data set included four classes of variables (ie, features):

1. Demographic variables: year of birth, sex, and age (calculated until January 1, 2015).
2. Numerical variables: A total of 13 continuous measurements are presented in the data set, including the number of MBS records, number of states, total amount of medical fees charged, total amount of medicare schedule fees, total amount of medical rebates paid, total number of MBS services, total duration of patients accessing medical services, number of PBS records, number of patient's PBS codes, total amount of medication cost paid by the government, total amount of medication cost self-paid, total number of prescriptions, and total duration of patients accessing PBS services.
3. Categorical variables: These are 3 relevant medications classified by the Anatomical Therapeutic Chemical code and patient state. The medications presented are drugs used in diabetes (code: A10), drugs used for the cardiovascular system and hypertension (code: C0), and lipid-modifying agents or drugs used for patients with high cholesterol (code: C10).
4. Date variables: The 4 date variables include the date of the first medical schedule, date of the last medical schedule, date of the first PBS claim, and date of the last PBS claim.

Among these variables, except for the year of birth, age, and numerical variables that were kept constant, other variables were transformed as follows: sex and medication variables were mapped into binary values, whereas patient state was converted into 6 binary variables corresponding to 6 states. The year of birth, date of first medical schedule, and date of first PBS claim were used to calculate the age at which the patient had the first medical schedule and the first PBS claim, respectively, and then removed. Regarding the prediction target variable, because PBS and MBS claim data on their own do not include information about patients' health outcomes, the labels must be inferred. Between the date of the last medical schedule and date of the last PBS claim, the latter was used to calculate the duration of patients discontinuing PBS and MBS services until January 1, 2015. Following this calculation, any patient who discontinued PBS and MBS for more than 180 days (6 months) was labeled *deceased*, otherwise *alive*. After preprocessing, the data set had 26 features and 1 label that used for model development.

In terms of feature scaling, each feature value was standardized to center around its mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation [34]. This step allows the algorithm to learn effectively as it eliminates

sensitivity to multiple features spanning varying degrees of magnitude, range, and units.

In terms of class distributions, there are only 93,164 patients out of the total number of 346,182 classified into the *deceased* group, whereas the rest are *alive* patients. This reflects a highly imbalanced class distribution, which might affect the learning performance of the infrequent class [35] because of the lack of samples. To address this issue, a separate experiment using SMOTE was conducted as a trial to enrich the training set.

Evaluation Metrics

Descriptive statistics were used to learn the characteristics of the study population, stratified by health outcome status (ie, alive or deceased). Models were derived from the training set and then assessed on the testing set by calculating the traditional accuracy, precision, and recall scores with the addition of brier loss. In addition, reporting discrimination and calibration is important for assessing a prediction model [36]. The area under the receiver operating characteristic curve (AUROC) score and the plotting reliability diagram (calibration curves) were also calculated to assess the performance of the AI models.

- Brier loss from scikit-learn measures the accuracy of probabilistic predictions by calculating the mean squared difference between the predicted probability assigned to the possible classes and the actual classes. It is composed of refinement loss and calibration loss so that the lower the Brier score is for a set of predictions, the better the predictions are calibrated or the better the model is.
- The AUROC score is used to measure the probability that the model ranks a random deceased patient higher than a random alive patient in terms of mortality rate. A higher AUROC score means that the model has a better ability to discriminate between deceased and alive populations.
- Calibration curve, a reliability diagram, is a line plot of the relative frequency of what was observed versus the predicted probability frequency. The closer the points appear along the main diagonal from the bottom left to the top right, the better calibrated a forecast or more reliable a model [37].

Hyperparameters

To develop the models, the study population was stratified into a training set, in which the mortality risk algorithms were derived, and a testing set, in which the algorithms were applied and tested. The training set consisted of 90.00% (311,564/346,182) of the study data set, and the testing set consisted of the remaining 10.00% (34,618/346,182). The training and testing sets were split at the patient level and in a stratifying manner according to class ratio so that patients did not appear in both the training and testing sets and the ratio of patient labels (*deceased* or *alive*) in both sets were equivalent to that of the study population. After stratified assignment, the hyperparameters were determined by using a grid search of 5-fold cross-validation to determine the values that led to the best accuracy. After the grid search, each algorithm was refitted to the training set with its best hyperparameters to derive the final models. Table 2 presents the parameter search space of the 4 algorithms and the grid results.

Table 2. Hyperparameters for grid search.

Algorithms and parameter name	Search space	Optimal
Logistic regression		
• Penalty	• (‘l1’, ‘l2’, ‘none’)	• l2
• C	• (0.01, 0.1, 1.0)	• 1.0
• tol	• (0.0001, 0.001, 0.01)	• 0.0001
• solver	• (‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’) (‘auto’, ‘ovr’, ‘multinomial’)	• lbfgs
• multi_class		• auto
Random forest		
• n_estimators	• (5, 10, 50, 100, 150)	• 100
• max_depth	• (1, 2, 3, 5, None)	• None
• max_features	• (’auto’, ’sqrt’)	• auto
• min_samples_splitmin_samples_leaf	• (2, 5, 10)	• 2
	• (1, 2, 4)	• 1
Extra trees		
• n_estimators	• (5, 10, 50, 100, 150)	• 100
• max_depth	• (1, 2, 3, 5, None)	• None
• max_features	• (’auto’, ’sqrt’)	• auto
• min_samples_splitmin_samples_leaf	• (2, 5, 10)	• 2
	• (1, 2, 4)	• 1
Gradient boosting trees		
• Loss	• (‘deviance’, ‘exponential’)	• deviance
• n_estimators	• (5, 10, 50, 100, 150)	• 100
• max_depth	• (1, 2, 3, 5)	• 3
• learning_rate	• (0.001, 0.01, 0.1)	• 0.1
• criterion	• (’friedman_mse’, ’mse’, ’mae’)	• friedman_mse

After the grid search, it was found that LR with L2 regularization, which is also known as Ridge Regression [38], produces the most accurate predictions in cross-validation, with the C value and tolerance rate of 1.0 and 0.0001, respectively. This can be explained by the fact that our data set had a small number of features, making L1 regularization, which is Lasso Regression and works well for feature selection in a data set with high dimensionality [39], less favorable. Next, both RF and ET achieved optimal accuracy after grid search with the max_depth *None* scheme. According to the scikit-learn team, in this scheme nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples, which is optimized at 2 in both cases. Besides, the number of trees grown in both algorithms is the same, 100 (n_estimators). Last, errors in GBT are minimized using the deviance loss function; there are also 100 trees built with the maximum number of nodes equal to 3.

To develop the DNN model, the study population was stratified into training and testing sets with ratios of 90% and 10%, respectively. The training set was then broken down into training and validation sets with the same ratio. The purpose of the validation set was to provide an unbiased evaluation of the model while tuning the weights of the model [40]. The input layer had 26 units corresponding to the number of features, whereas the output layer had one unit. At the last step, sigmoid was used as the activation function to return the sigmoid values of the final output. The architecture of the DNN used is composed of 3 fully connected hidden layers. The number of neurons in each hidden layer are 128, 64 and 32, respectively, and the rectified linear unit is used as the activation function. During the training process, the parameters of the DNN are initialized using uniform initialization [41]. For each batch of training data, parameters of the DNN were modified gradually to decrease the cross-entropy of the loss function. A callback was set to stop the training process after 10 epochs when the model reached the highest value of AUROC.

After the training process, all models were evaluated using the holdout (10%) testing set. The final results were compared and used to make recommendations.

Model Performance

In our experiments, we trained the models using the original learning sample and then applied SMOTE to further improve their performance.

Performance Without SMOTE

The details of the model performance without SMOTE are presented in Table 3. After adjusting for multiple comparisons, there was no significant difference in accuracy among RF (98.5%), GBT (98.4%), LR (97.8%), ET (97.9%), and DNN (97.1%). In terms of discrimination, GBT and RF achieved the highest AUROC (97.8% and 97.7%, respectively), followed by LR and ET (96.4% and 96.8%, respectively), whereas DNN was the least discriminative (95.3%). In terms of brier loss, GBT and RF produced the smallest difference between the probability assigned to the predicted classes and the probability of the actual class (both 0.012), whereas DNN predictions showed the largest difference (0.024).

Table 3. Performance metrics of machine learning models without the Synthetic Minority Oversampling Technique.

Algorithms	Accuracy	Area under the receiver operating characteristic curve	Precision	Recall	Brier loss
Logistic regression	97.8	96.4	98.5[a]	93.4	0.016
Random forest	98.5[b]	97.7	98.1	96.1	0.012[c]
Extra trees	97.9	96.8	98.1	94.2	0.016
Gradient boosting trees	98.4	97.8[d]	97.5	96.5[e]	0.012[c]
Artificial neural network	97.1	95.3	96.6	91.8	0.024

[a]The highest precision.

[b]The highest accuracy.

[c]The least Brier loss.

[d]The highest area under the receiver operating characteristic curve.

[e]The highest recall.

According to Table 4 showing the training times, LR turns out to be superior compared with other models with less than 1-min training time. However, DNN takes up to 30 minutes to train. This could be explained by the complexity level of the 2 algorithms; whereas LR is a very simple and straightforward model based on a linear regression equation, DNN is an architecture that is composed of many neurons, layers, and more complex activation functions.

Clearly, all of our models show very similar behavior for the 2 classes (Figures 6-10). According to the confusion matrices, RF and GBT managed to identify the *deceased* patients with higher accuracy than other algorithms. Compared with other models, DNN classifies a larger number of *deceased* patients as *alive*.

Table 4. Training time of machine learning models without Synthetic Minority Oversampling Technique.

Algorithms	Training time (seconds)
Logistic regression	6.6[a]
Random forest	106.8
Extra trees	46.8
Gradient boosting trees	186
Artificial neural network	1277.4

[a]The least training time.

Figure 6. Confusion matrices of logistic regression.

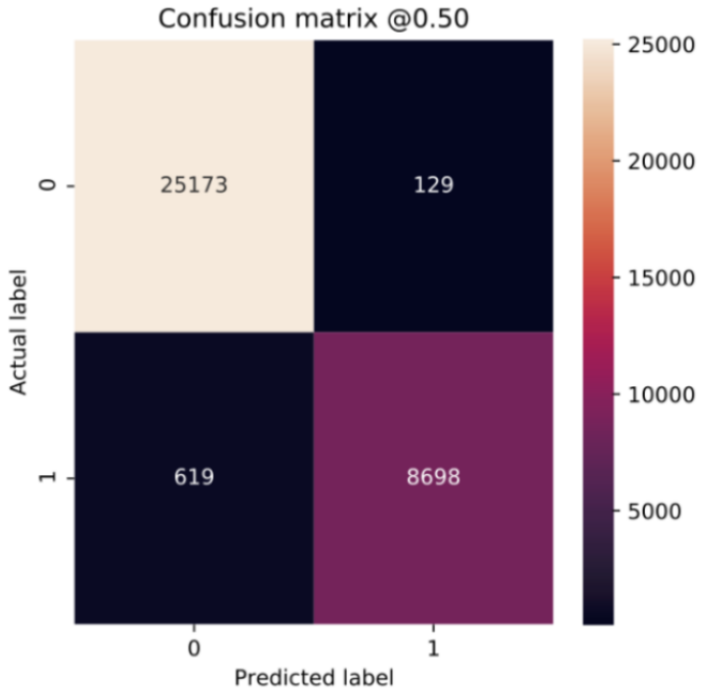


Figure 7. Confusion matrix of random forest.

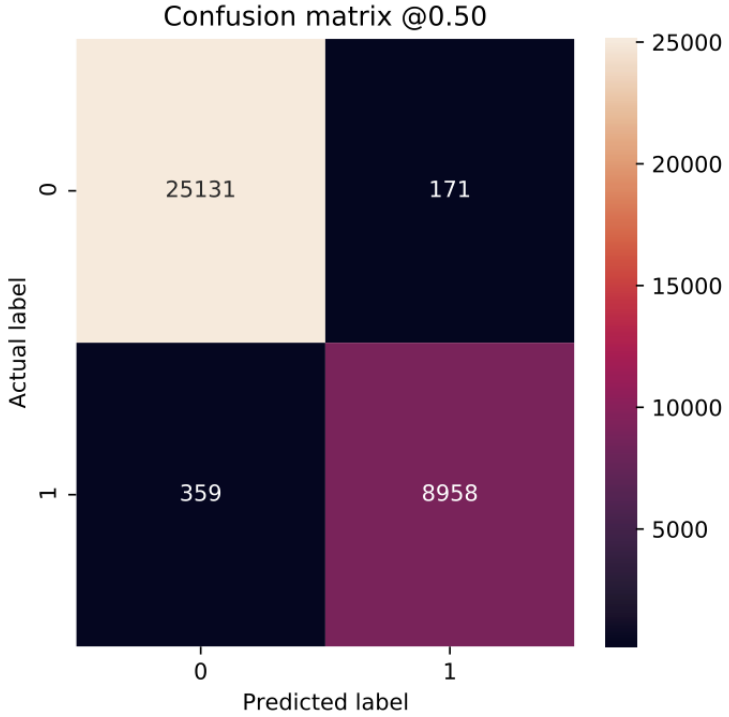


Figure 8. Confusion matrix of extra trees.

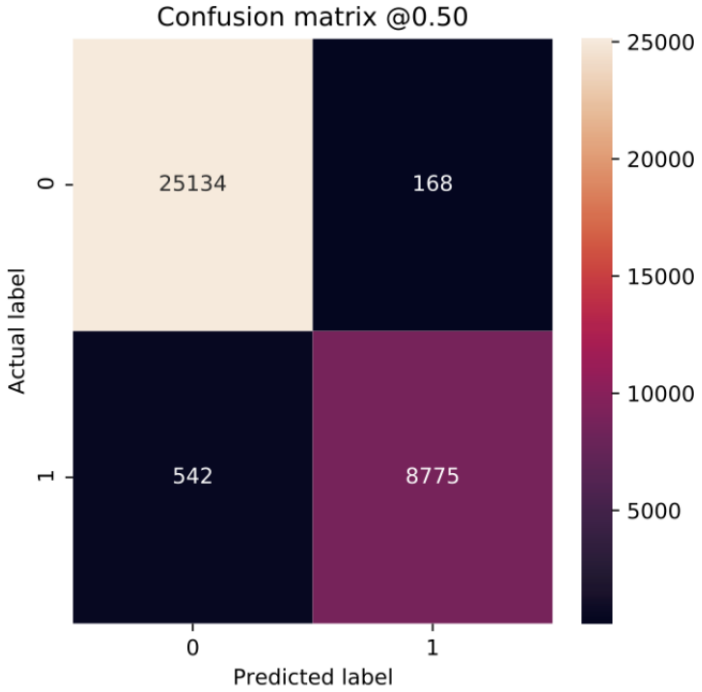


Figure 9. Confusion matrix of gradient boosting trees.

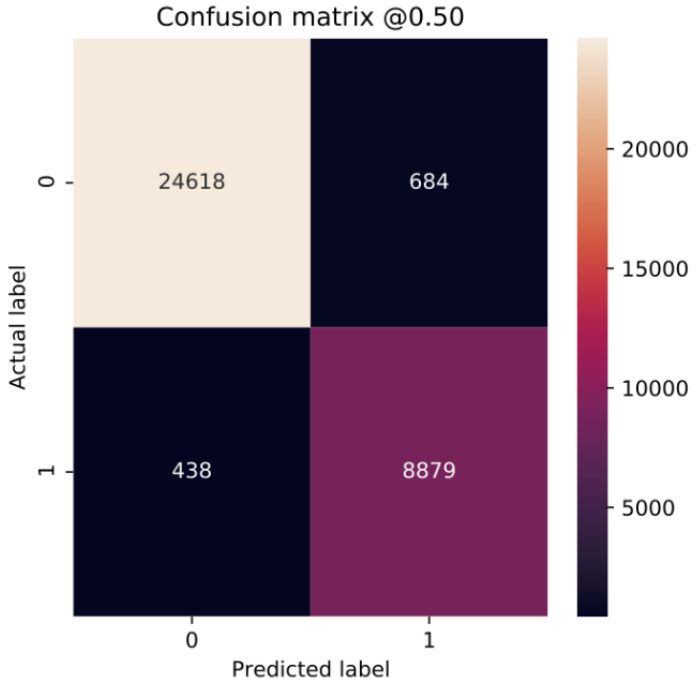
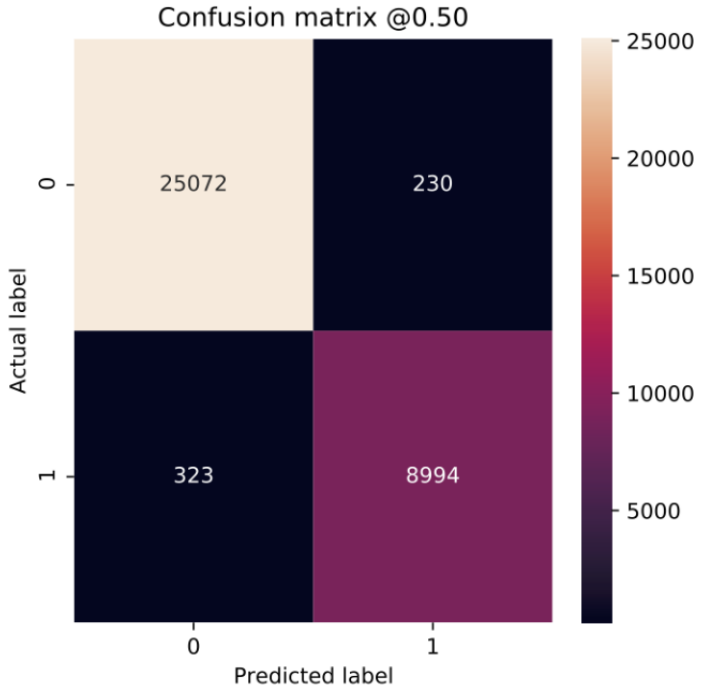


Figure 10. Confusion matrix of artificial neural network.



In terms of prediction reliability, calibration curves for the 5 models in Figures 11-20 show that LG was the least calibrated compared with the other 4 algorithms, highly overestimating patient death risks at all levels of probabilities. RF was well calibrated for patients with a lower mortality rate and overestimated the risk of death when the probability of risk was more than 50%. ET's goodness of fit was only seen in the probability of death at 30%, whereas it underestimated and overestimated the risk for patients with lower and higher probabilities of death, respectively. Predictions by GBT and DNN were the most well calibrated, whereas DNN slightly overestimated patients with probabilities of death greater than 10% and less than 90%.

Figure 11. Calibration curve of random forest without Synthetic Minority Oversampling Technique.

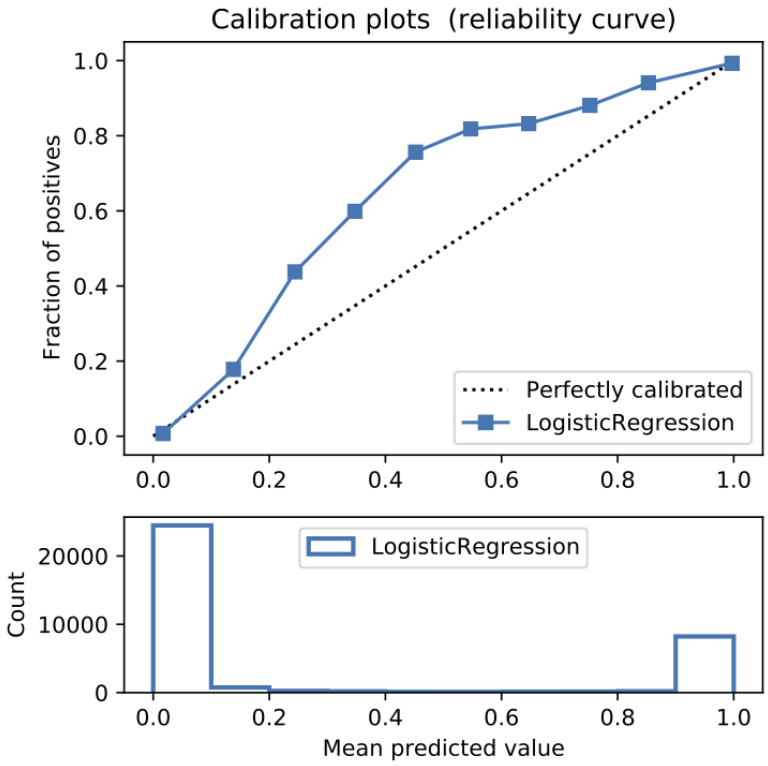


Figure 12. Calibration curve of random forest without Synthetic Minority Oversampling Technique.

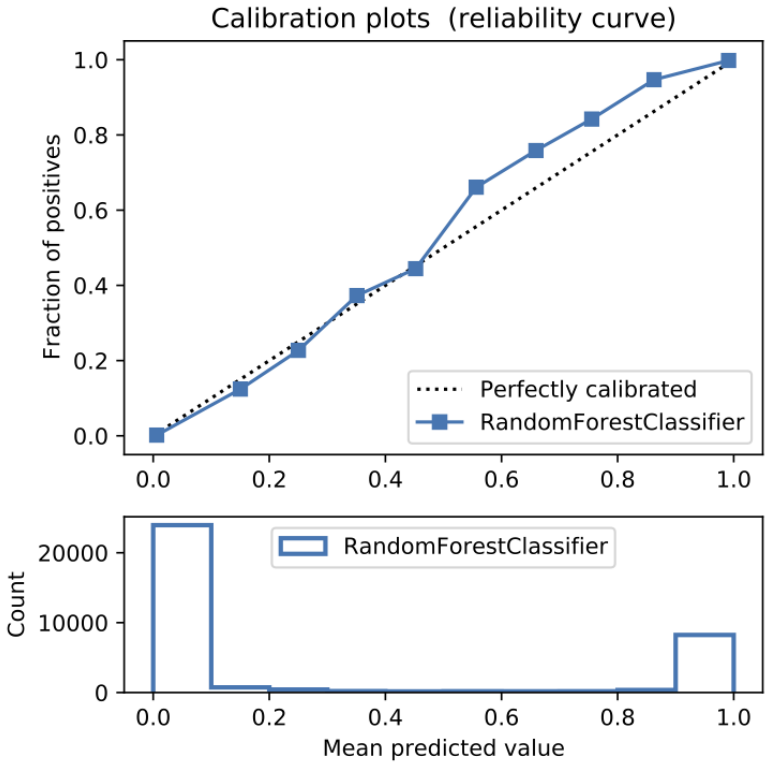


Figure 13. Calibration curve of extra trees without Synthetic Minority Oversampling Technique.

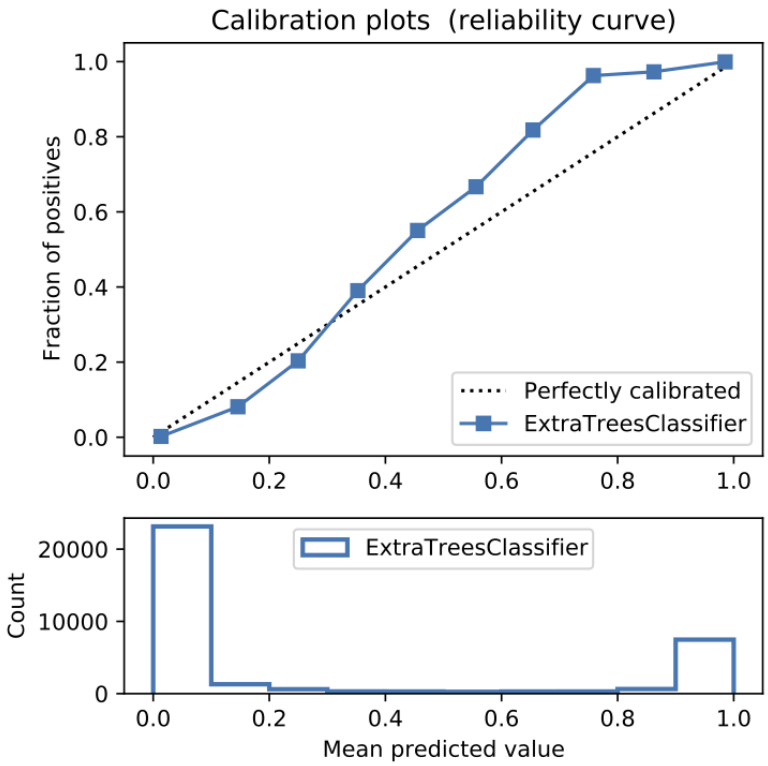


Figure 14. Calibration curve of gradient boosting trees without Synthetic Minority Oversampling Technique.

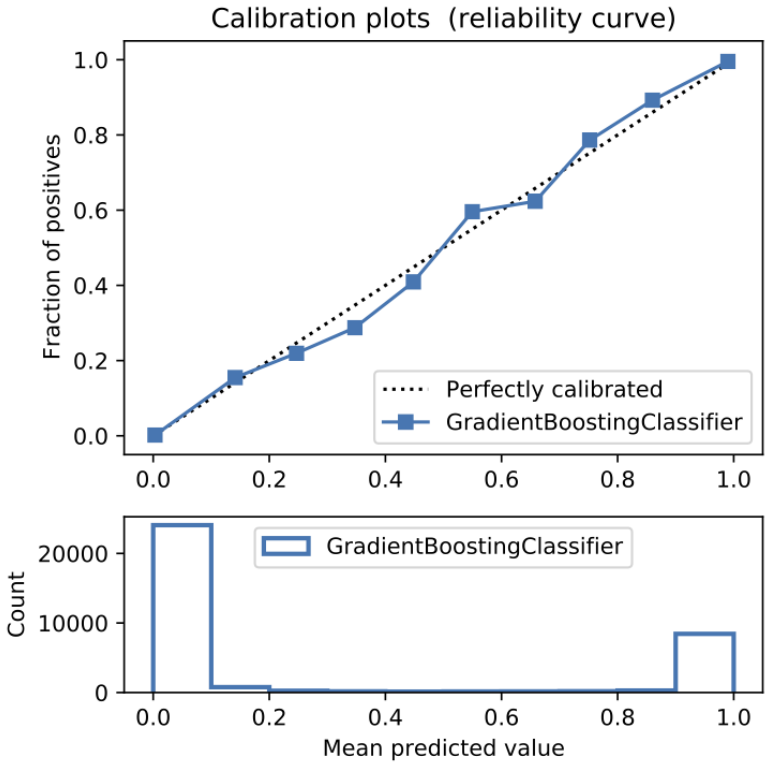


Figure 15. Calibration curve of artificial neural network without Synthetic Minority Oversampling Technique.

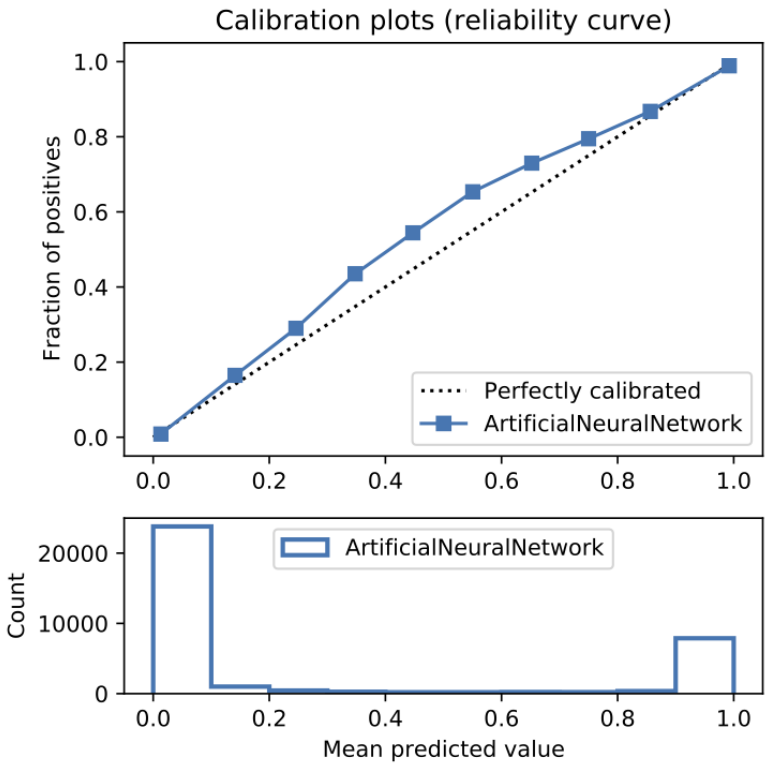


Figure 16. Calibration curve of logistic regression with Synthetic Minority Oversampling Technique.

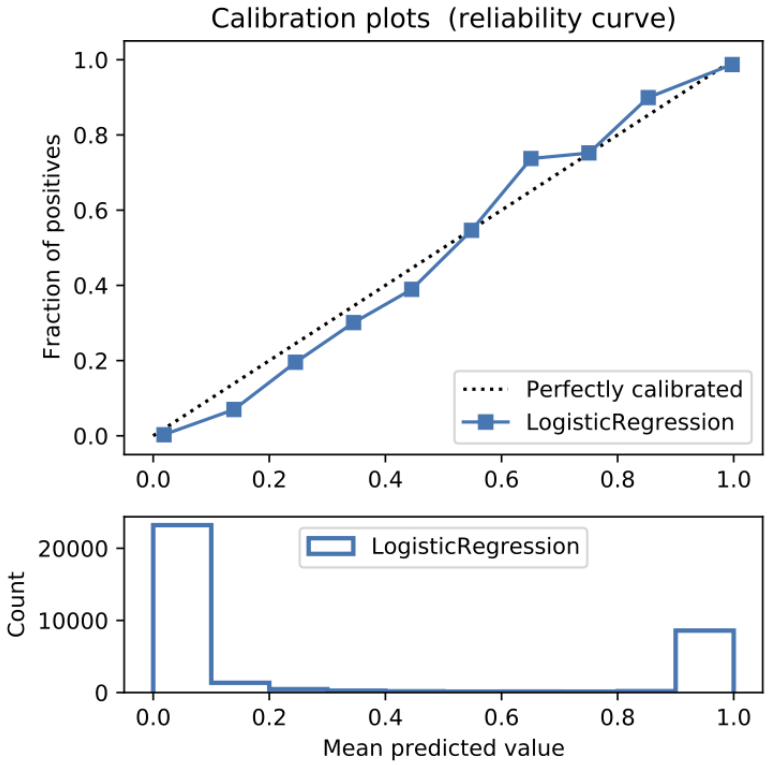


Figure 17. Calibration curve of random forest with Synthetic Minority Oversampling Technique.

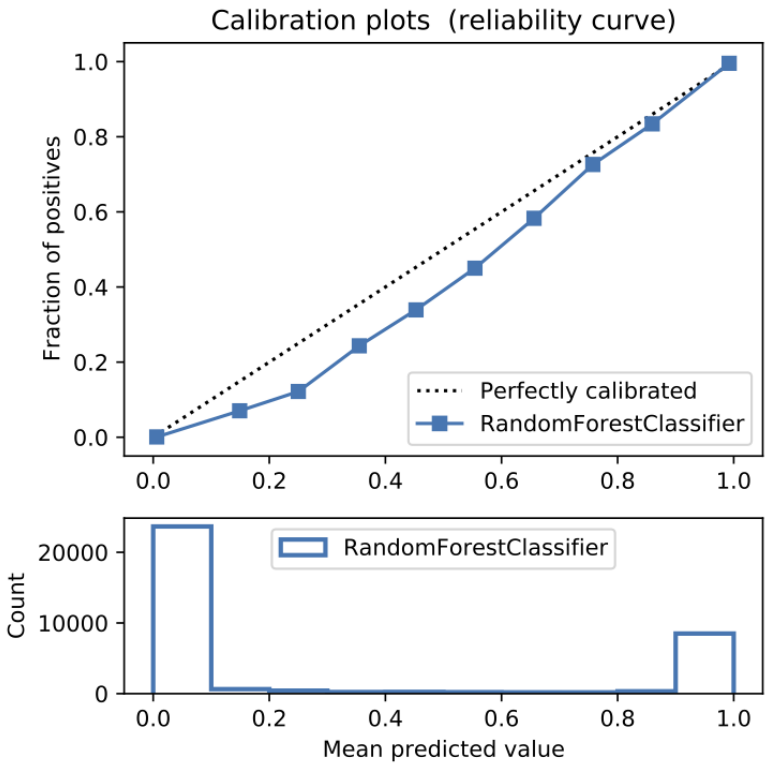


Figure 18. Calibration curve of extra trees with Synthetic Minority Oversampling Technique.

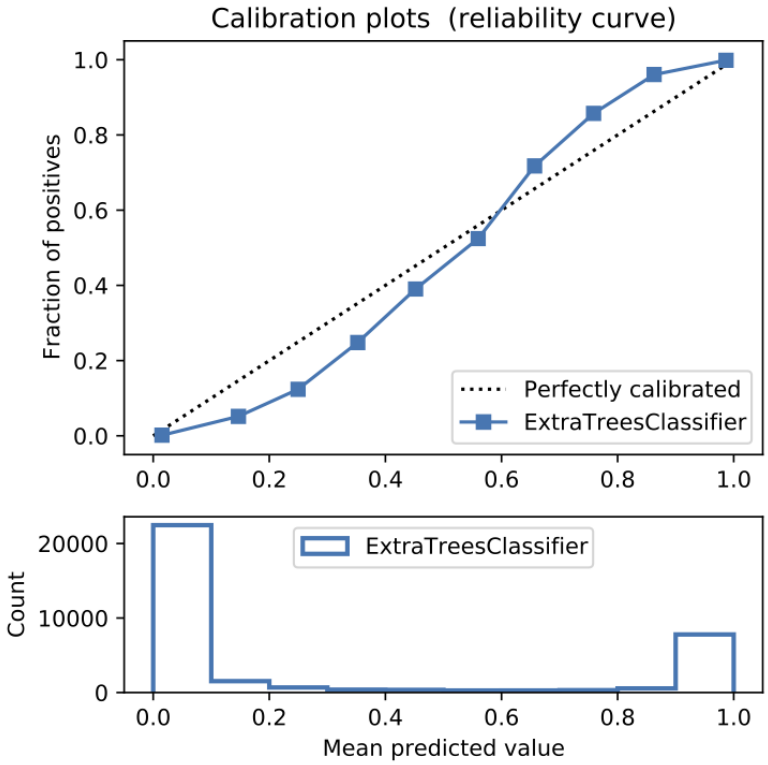


Figure 19. Calibration curve of gradient boosting trees with Synthetic Minority Oversampling Technique.

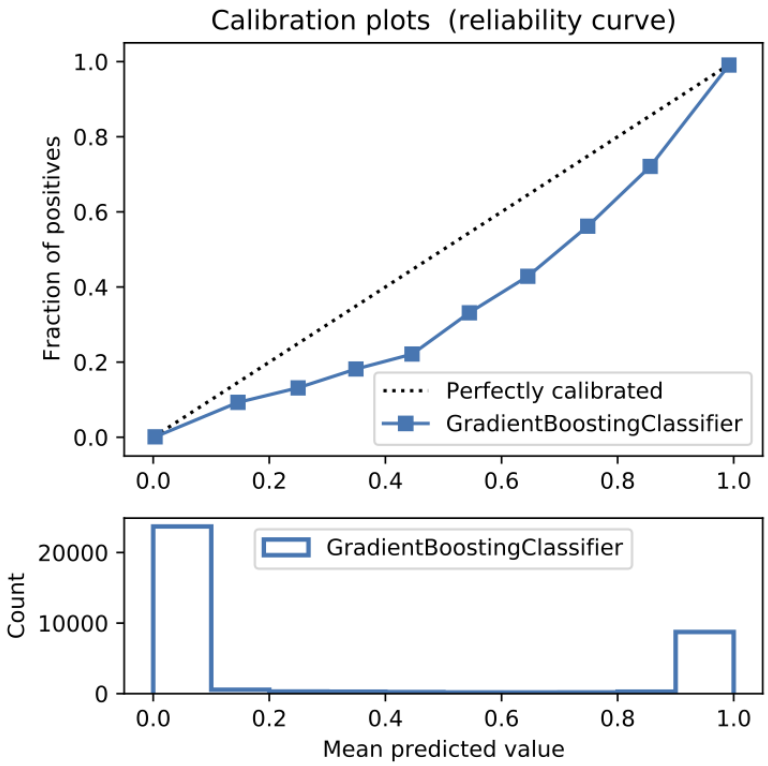
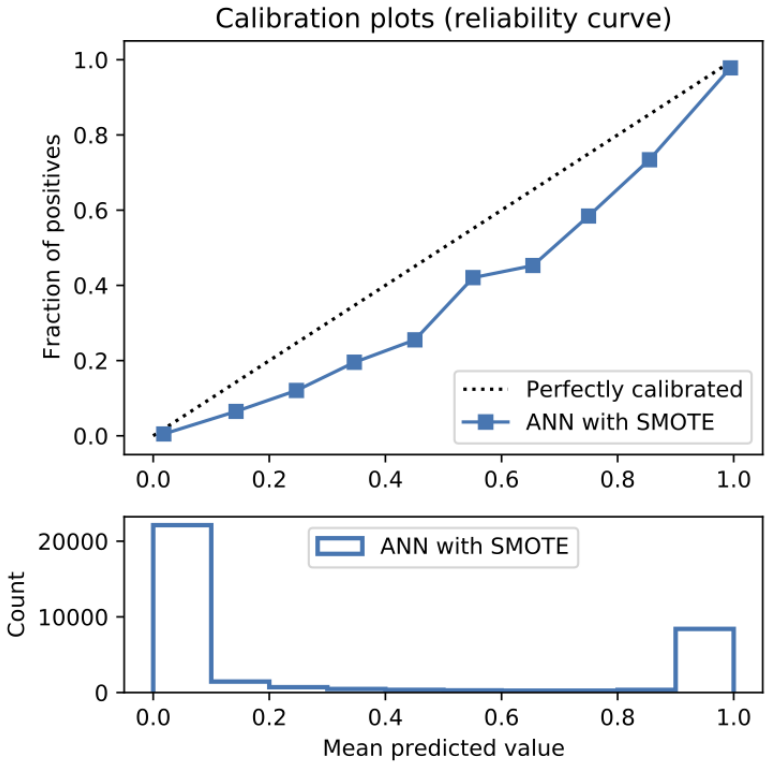


Figure 20. Calibration curve of artificial neural network with Synthetic Minority Oversampling Technique.



Performance With SMOTE

Details of the model performance with SMOTE are presented in Table 5, and their calibration plots are displayed in Figure 8. As can be seen in Table 5, SMOTE slightly improves the performance (in italics) of the 5 models. However, it helps calibrate the predictions of LR significantly. After upsampling, the LR model no longer overestimates death risks of the patient, and its predictions are more closely aligned with the perfectly calibrated line. Meanwhile, ET is now seen as having goodness of fit in predictions of patients with death risk between 50% and 60% but still underestimates and overestimates those with low and high death risks, respectively. On the other hand, RF predictions change from being well calibrated for less than 50% probabilities of death risk and overestimating higher ones into being well calibrated for greater than 80% probabilities of death risk and underestimating the rest. More interestingly, DNN and GBT are subject to adversarial effects from the upsampling technique, generally underestimate the risk.

Table 5. Performance metrics of machine learning models with the Synthetic Minority Oversampling Technique.

Algorithms	Accuracy	Area under the receiver operating characteristic curve	Precision	Recall	Brier loss
Logistic regression	98.2	97.4	97.3[a]	95.9	0.015
Random forest	98.4[b]	98.0[c]	96.8	97.3	0.012[d]
Extra trees	98.1	97.4	97.1	95.8	0.016
Gradient boosting trees	98.1	97.9	95.2	97.7[e]	0.014
Artificial neural network	96.7	96.2	93.0	95.1	0.026

[a]The highest precision.

[b]The highest accuracy.

[c]The highest area under the receiver operating characteristic curve.

[d]The least Brier loss.

[e]The highest recall.

In short, SMOTE is only helpful for further improving the model performance and prediction calibration of LG. Meanwhile, using or not using SMOTE does not affect the performance of RF and ET in predicting mortality in patients with CVD. Last, SMOTE introduces an adversarial effect into the GBT and DNN models, making their predictions less reliable, and these 2 models already work well with class imbalanced data.

In terms of training duration, as shown in Table 6, using SMOTE requires more computing time for all the algorithms. However, LR is still the most time-efficient model even when applying

SMOTE and produces higher accuracy and better prediction performance in terms of AUROC, recall, and brier loss compared with LR with original data. Furthermore, SMOTE helps LR outperform ET and become the second-best algorithm after RF. Clearly, when introducing SMOTE into the table, ET and LR are worth considering for this data set.

Table 6. Training time of machine learning models with the Synthetic Minority Oversampling Technique.

Algorithms	Training time (seconds)
Logistic regression	292.9[a]
Random forest	497.9
Extra trees	347.5
Gradient boosting trees	648.1
Artificial neural network	5480.3

[a]The least training time.

Discussion

Principal Findings

This study shows that structured medical and pharmaceutical claims data can be used as input for AI models to accurately predict the mortality risk of individuals with CVD. The LR, RF, ET, GBT, and ANN models trained in this study had high accuracy (ie, 97.0%-98.0%) and discrimination (ie, AUROC 95.0%-98.0%) in predicting the mortality rate, which are much higher than for traditional statistical models such as the Cox Proportional-Hazards model [42] or the models trained with traditional electrical health records [43-45].

Although there was no statistically significant difference in accuracy among the 5 experimental algorithms, the RF model had an advantage over the other models. In addition, the RF model outperformed the other models in terms of recall and brier loss. In terms of discrimination and calibration, the GBT proved to be the most superior. Without SMOTE, LR is unable to make highly calibrated predictions while using SMOTE significantly improves the reliability of the model's predictions. All models with SMOTE had very high precision (ie, 93.0%-97.0%) and recall (ie, 95.0%-97.0%), particularly when compared with other LR and RF prognostic models that did not deal with class imbalance published in the literature [44,45]. On the other hand, although the ANN had the most moderate performance among the experimental algorithms, it was proven to be efficient even with class imbalanced data. It is also suggested that ANNs are capable of predicting CVD mortality rates more accurately than other ML algorithms if more feature-engineering techniques are applied [46,47], indicating it is a very promising area for further research.

To our knowledge, this is the first study comparing AI algorithms using medical and pharmaceutical claims data to predict mortality in a large general cardiology population. Unlike previously developed ML-based prognostic tools in cardiology that used the clinical information of patients, including clinical features [43-45], our models were trained only on claims data of patients with CVD. These claims data primarily provide information about a patient's medical scheduling and pharmaceutical dispensing history, which reflect the patient's disease treatment cost, access patterns, and medications but not the patient's state of health or other clinical indices. Furthermore, compared with previously published classifiers in cardiology, our models used fewer features and are comparatively more efficient than previously trained models in the general cardiology setting.

Limitations

Despite high accuracy and strong discrimination, some models, including RF, ET, and ANN, still have not yielded optimal calibrations. This means that the distribution and behavior of the predicted probability is not similar to the distribution and behavior of the probability observed in training data. To increase the reliability of AI algorithms, other techniques should be investigated to better calibrate and improve the performance of these models, especially ANNs.

Conclusions

We developed, validated, and compared 5 AI architectures to predict the mortality rate of patients with CVD. On the basis of the evaluation results, we can draw the following conclusions or insights that could help with the choice of AI models: (1) without health indices or health condition information, AI architectures are able to accurately predict mortality of patients with CVD before a clinic visit using only medical scheduling and pharmaceutical dispensing claims data; (2) although there was no statistically significant difference in accuracy among the experimental AI algorithms, the tree-based, that is, RF and GBT models have an advantage compared with other models; (3) although the regression-based LR method produces predictions having the least calibration level because of a lack of minority class samples, the upsampling technique, that is, SMOTE helps significantly improve the reliability of this algorithm's predictions; and (iv) tree-based algorithms and densely connected neural networks perform well with class imbalanced data. Finally, this study showed the feasibility and effectiveness of different AI architectures based on structured medical scheduling and pharmaceutical dispensing claims data in identifying patients with CVD who had a risk of mortality; AI algorithms can be a useful tool for precise decision making. Future research, considering the promising potential of the ANN, should focus on improving the prediction performance of this algorithm. It is suggested that ANNs are capable of predicting CVD mortality rates more accurately than other ML algorithms if more feature-engineering techniques are applied, indicating they are a very promising area for further research.

Acknowledgments

The authors would like to thank Dr Dennis Wollersheim and Dr Shaun Purkiss from La Trobe University for helping them prepare the data set used in this research.

Conflicts of Interest

None declared.

References

1. AIHW. Cardiovascular disease: Australian facts 2011. In: AIHW; Cat. no: CVD 53. Australia: AIHW; 2011:53.
2. Aguilar-Palacio I, Malo S, Lallana M, Feja C, González J, Moreno-Franco B, et al. Co-prescription patterns of cardiovascular preventive treatments: a cross-sectional study in the Aragon worker' health study (Spain). BMJ Open 2019 Apr 14;9(4):e023571 [FREE Full text] [doi: 10.1136/bmjopen-2018-023571] [Medline: 30987984]
3. Stoeldraijer L, van Duin C, van Wissen L, Janssen F. Impact of different mortality forecasting methods and explicit assumptions on projected future life expectancy: the case of the Netherlands. DemRes 2013 Aug 27;29:323-354. [doi: 10.4054/demres.2013.29.13]
4. Booth H, Tickle L. Mortality modelling and forecasting: a review of methods. Ann Actuar Sci 2011 May 10;3(1-2):3-43. [doi: 10.1017/s1748499500000440]
5. Mortazavi BJ, Bucholz EM, Desai NR, Huang C, Curtis JP, Masoudi FA, et al. Comparison of machine learning methods with national cardiovascular data registry models for prediction of risk of bleeding after percutaneous coronary intervention. JAMA Netw Open 2019 Jul 03;2(7):e196835 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.6835] [Medline: 31290991]
6. Cleophas TJ, Zwinderman AH. Machine learning and unsolved questions. In: Machine Learning in Medicine. Dordrecht: Springer; 2013:205-214.
7. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One 2017 Apr 4;12(4):e0174944 [FREE Full text] [doi: 10.1371/journal.pone.0174944] [Medline: 28376093]
8. Su M, Miften M, Whiddon C, Sun X, Light K, Marks L. An artificial neural network for predicting the incidence of radiation pneumonitis. Med Phys 2005 Feb 12;32(2):318-325. [doi: 10.1118/1.1835611] [Medline: 15789575]
9. Wang Z, Zhu Y, Li D, Yin Y, Zhang J. Feature rearrangement based deep learning system for predicting heart failure mortality. Comput Methods Programs Biomed 2020 Jul;191:105383. [doi: 10.1016/j.cmpb.2020.105383] [Medline: 32062185]
10. Hung CY, Chen WC, Lal PT, Lee CC. Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database. In: Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2017 Presented at: Comparing deep neural network and other machine learning algorithms for stroke prediction in a large-scale population-based electronic medical claims database, inth Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 11-15, 2017; Jeju, Korea (South) p. 3110-3113. [doi: 10.1109/embc.2017.8037515]
11. Nirschl JJ, Janowczyk A, Peyster EG, Frank R, Margulies KB, Feldman MD, et al. PLoS One 2018;13(4):e0192726 [FREE Full text] [doi: 10.1371/journal.pone.0192726] [Medline: 29614076]
12. Martin-Isla C, Campello VM, Izquierdo C, Raisi-Estabragh Z, Baebler B, Petersen SE, et al. Image-based cardiac diagnosis with machine learning: a review. Front Cardiovasc Med 2020 Jan 24;7:1 [FREE Full text] [doi: 10.3389/fcvm.2020.00001] [Medline: 32039241]
13. Kilic A. Artificial intelligence and machine learning in cardiovascular health care. Ann Thorac Surg 2020 May;109(5):1323-1329. [doi: 10.1016/j.athoracsur.2019.09.042] [Medline: 31706869]
14. Small AM, Kiss DH, Zlatsin Y, Birtwell DL, Williams H, Guerraty MA, et al. Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease. J Biomed Inform 2017 Aug;72:77-84 [FREE Full text] [doi: 10.1016/j.jbi.2017.06.016] [Medline: 28624641]
15. Chu J, Dong W, He K, Duan H, Huang Z. Using neural attention networks to detect adverse medical events from electronic health records. J Biomed Inform 2018 Nov;87:118-130 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.002] [Medline: 30336262]
16. Matheny ME, Ohno-Machado L, Resnic F. Discrimination and calibration of mortality risk prediction models in interventional cardiology. J Biomed Inform 2005 Oct;38(5):367-375 [FREE Full text] [doi: 10.1016/j.jbi.2005.02.007] [Medline: 16198996]
17. Matheny ME, Resnic FS, Arora N, Ohno-Machado L. Effects of SVM parameter optimization on discrimination and calibration for post-procedural PCI mortality. J Biomed Inform 2007 Dec;40(6):688-697 [FREE Full text] [doi: 10.1016/j.jbi.2007.05.008] [Medline: 17600771]

18. Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. J Biomed Inform 2016 Apr;60:260-269 [FREE Full text] [doi: 10.1016/j.jbi.2016.01.009] [Medline: 26844760]
19. Cleophas TJ, Zwinderman AH. Machine learning and unsolved questions. In: Machine Learning in Medicine. Dordrecht: Springer; 2013:205-214.
20. Roth C, Payne PR, Weier RC, Shoben AB, Fletcher EN, Lai AM, et al. The geographic distribution of cardiovascular health in the stroke prevention in healthcare delivery environments (SPHERE) study. J Biomed Inform 2016 Apr;60:95-103 [FREE Full text] [doi: 10.1016/j.jbi.2016.01.013] [Medline: 26828957]
21. Paterick TE, Patel N, Tajik AJ, Chandrasekaran K. Improving health outcomes through patient education and partnerships with patients. Proc (Bayl Univ Med Cent) 2017 Jan 11;30(1):112-113 [FREE Full text] [doi: 10.1080/08998280.2017.11929552] [Medline: 28152110]
22. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression, third Edition. In: Wiley Online Library. Hoboken, New Jersey: Wiley Online Library; 2013.
23. Friedl MA, Brodley CE. Decision tree classification of land cover from remotely sensed data. Remote Sens Environ 1997 Sep;61(3):399-409. [doi: 10.1016/s0034-4257(97)00049-7]
24. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019 Dec 21;19(1):281 [FREE Full text] [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]
25. Quinlan JR. Induction of decision trees. Mach Learn 1986 Mar;1(1):81-106. [doi: 10.1007/bf00116251]
26. Breiman L. Machine learning. Random forests 2001;1(5):45. [doi: 10.1201/9780367816377-11]
27. John V, Liu Z, Guo C, Mita S, Kidono K. Real-time lane estimation using deep features and extra trees regression. Image and Video Technology 2015:733. [doi: 10.1007/978-3-319-29451-3_57]
28. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006 Mar 2;63(1):3-42. [doi: 10.1007/s10994-006-6226-1]
29. Friedman JH. Stochastic gradient boosting. Computa Stati and Data Anal 2002 Feb;38(4):367-378. [doi: 10.1016/s0167-9473(01)00065-2]
30. Rahman S, Irfan M, Raza M, Moyeezullah Ghori K, Yaqoob S, Awais M. Performance analysis of boosting classifiers in recognizing activities of daily living. Int J Environ Res Public Health 2020 Feb 08;17(3):1082 [FREE Full text] [doi: 10.3390/ijerph17031082] [Medline: 32046302]
31. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. 1943. Bull Math Biol 1990;52(1-2):99-115. [Medline: 2185863]
32. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986 Oct;323(6088):533-536. [doi: 10.1038/323533a0]
33. Reid R. From functional architecture to functional connectomics. Neuron 2012 Jul 26;75(2):209-217 [FREE Full text] [doi: 10.1016/j.neuron.2012.06.031] [Medline: 22841307]
34. sklearn.preprocessing.StandardScaler. scikit-learn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html [accessed 2021-02-04]
35. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. IDA 2002 Nov 15;6(5):429-449. [doi: 10.3233/IDA-2002-6504]
36. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models. Epidemiology 2010;21(1):128-138. [doi: 10.1097/ede.0b013e3181c30fb2]
37. Brownlee J. How and when to use a calibrated classification model with scikit-learn. URL: https://machinelearningmastery.com/calibrated-classification-model-in-scikit-learn/ [accessed 2021-02-23]
38. Hoerl AE, Kannard RW, Baldwin KF. Ridge regression:some simulations. Commu Stat 1975 Jan;4(2):105-123. [doi: 10.1080/03610927508827232]
39. Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc 2018 Dec 05;58(1):267-288. [doi: 10.1111/j.2517-6161.1996.tb02080.x]
40. Brownlee J. What is the difference between test and validation datasets. URL: https://machinelearningmastery.com/difference-test-validation-datasets/ [accessed 2021-02-23]
41. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. 2015 Presented at: IEEE international conference on computer vision; 2015; Corfu, Greece p. 1026-1034. [doi: 10.1109/iccv.2015.123]
42. Hata J, Nagai A, Hirata M, Kamatani Y, Tamakoshi A, Yamagata Z, Biobank Japan Cooperative Hospital Group, et al. Risk prediction models for mortality in patients with cardiovascular disease: the BioBank Japan project. J Epidemiol 2017 Mar;27(3S):71-76 [FREE Full text] [doi: 10.1016/j.je.2016.10.007] [Medline: 28142037]
43. Kwon J, Kim K, Jeon K, Lee SE, Lee H, Cho H, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One 2019 Jul 8;14(7):e0219302 [FREE Full text] [doi: 10.1371/journal.pone.0219302] [Medline: 31283783]

44. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 2020 Feb 03;20(1):16 [FREE Full text] [doi: 10.1186/s12911-020-1023-5] [Medline: 32013925]
45. Lee DS, Austin PC, Rouleau JL, Liu PP, Naimark D, Tu JV. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. J Am Med Assoc 2003 Nov 19;290(19):2581-2587. [doi: 10.1001/jama.290.19.2581] [Medline: 14625335]
46. Wang Z, Zhu Y, Li D, Yin Y, Zhang J. Feature rearrangement based deep learning system for predicting heart failure mortality. Comput Methods Programs Biomed 2020 Jul;191:105383. [doi: 10.1016/j.cmpb.2020.105383] [Medline: 32062185]
47. Sadati N, Nezhad MZ, Chinnam RB, Zhu D. Representation learning with autoencoders for electronic health records: a comparative study. URL: https://arxiv.org/abs/1908.09174 [accessed 2021-02-23]

Abbreviations

ANN: artificial neural network
AUROC: area under the receiver operating characteristic curve
CVD: cardiovascular disease
EHR: electronic health record
ET: extra trees
GBT: gradient boosting trees
LR: logistic regression
MBS: Medicare Benefits Schedule
ML: machine learning
PBS: Pharmaceutical Benefits Scheme
RF: random forest
SMOTE: Synthetic Minority Oversampling Technique

Edited by C Lovis; submitted 14.10.20; peer-reviewed by W Zhang, X Chang; comments to author 07.11.20; revised version received 17.11.20; accepted 05.12.20; published 01.04.21.

Please cite as:
Tran L, Chi L, Bonti A, Abdelrazek M, Chen YPP
Mortality Prediction of Patients With Cardiovascular Disease Using Medical Claims Data Under Artificial Intelligence Architectures: Validation Study
JMIR Med Inform 2021;9(4):e25000
URL: https://medinform.jmir.org/2021/4/e25000
doi:10.2196/25000
PMID:33792549

Original Paper

TOP-Net Prediction Model Using Bidirectional Long Short-term Memory and Medical-Grade Wearable Multisensor System for Tachycardia Onset: Algorithm Development Study

Xiaoli Liu[1*], BSc; Tongbo Liu[2*], BSc; Zhengbo Zhang[3,4*], DPhil; Po-Chih Kuo[5], DPhil; Haoran Xu[6], BSc; Zhicheng Yang[7], DPhil; Ke Lan[8], MSc; Peiyao Li[9], MSc; Zhenchao Ouyang[10], DPhil; Yeuk Lam Ng[11], BSc; Wei Yan[12], MD; Deyu Li[1], DPhil

[1]Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China

[2]Department of Computer Management and Application, Chinese PLA General Hospital, Beijing, China

[3]Center for Artificial Intelligence in Medicine, Chinese PLA General Hospital, Beijing, China

[4]Department of Biomedical Engineering, Chinese PLA General Hospital, Beijing, China

[5]Laboratory for Computational Physiology, Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, MA, United States

[6]Medical School of Chinese PLA, Beijing, China

[7]US Research Lab, PingAn Tech, San Francisco, CA, United States

[8]Beijing SensEcho Science & Technology Co., Ltd, Beijing, China

[9]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[10]Hangzhou Innovation Institute, Beihang University, Beijing, China

[11]Faculty of Arts & Science, University of Toronto, Toronto, ON, Canada

[12]Department of Hyperbaric Oxygen, Chinese PLA General Hospital, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Deyu Li, DPhil
Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering
Beihang University
No. 37 Xueyuan Road, Haidian District
Beijing, 100083
China
Phone: 86 010 82339093
Email: deyuli@buaa.edu.cn

Abstract

Background: Without timely diagnosis and treatment, tachycardia, also called tachyarrhythmia, can cause serious complications such as heart failure, cardiac arrest, and even death. The predictive performance of conventional clinical diagnostic procedures needs improvement in order to assist physicians in detecting risk early on.

Objective: We aimed to develop a deep tachycardia onset prediction (TOP-Net) model based on deep learning (ie, bidirectional long short-term memory) for early tachycardia diagnosis with easily accessible data.

Methods: TOP-Net leverages 2 easily accessible data sources: vital signs, including heart rate, respiratory rate, and blood oxygen saturation (SpO_2) acquired continuously by wearable embedded systems, and electronic health records, containing age, gender, admission type, first care unit, and cardiovascular disease history. The model was trained with a large data set from an intensive care unit and then transferred to a real-world scenario in the general ward. In this study, 3 experiments incorporated merging patients' personal information, temporal memory, and different feature combinations. Six metrics (area under the receiver operating characteristic curve [AUROC], sensitivity, specificity, accuracy, F1 score, and precision) were used to evaluate predictive performance.

Results: TOP-Net outperformed the baseline models on the large critical care data set (AUROC 0.796, 95% CI 0.768-0.824; sensitivity 0.753, 95% CI 0.663-0.793; specificity 0.720, 95% CI 0.645-0.758; accuracy 0.721; F1 score 0.718; precision 0.686) when predicting tachycardia onset 6 hours in advance. When predicting tachycardia onset 2 hours in advance with data acquired from our hospital using the transferred TOP-Net, the 6 metrics were 0.965, 0.955, 0.881, 0.937, 0.793, and 0.680, respectively. The best performance was achieved using comprehensive vital signs (heart rate, respiratory rate, and SpO_2) statistical information.

Conclusions: TOP-Net is an early tachycardia prediction model that uses 8 types of data from wearable sensors and electronic health records. When validated in clinical scenarios, the model achieved a prediction performance that outperformed baseline models 0 to 6 hours before tachycardia onset in the intensive care unit and 2 hours before tachycardia onset in the general ward. Because of the model's implementation and use of easily accessible data from wearable sensors, the model can assist physicians with early discovery of patients at risk in general wards and houses.

(JMIR Med Inform 2021;9(4):e18803) doi:10.2196/18803

KEYWORDS

tachycardia onset; early prediction; deep neural network; wearable monitoring system; electronic health record

Introduction

Tachycardia, a heart rhythm disorder, is defined as an adult resting heart rate that exceeds 100 bpm [1]. According to the mechanisms, causes, expressions and outcomes, tachycardia can be classified as sinus tachycardia, atrial fibrillation, atrial flutter, ventricular tachycardia, or ventricular fibrillation [2]. Spontaneous ventricular tachyarrhythmia is a major cause of sudden cardiac death; approximately 180,000 to 300,000 people suffer from this condition in the US yearly [3,4]. Atrial fibrillation is a risk factor for stroke, congestive heart failure, and premature death. Patients suffering from atrial fibrillation for the first time have a high rate of mortality [5,6]. In addition, tachycardia has been correlated to poor outcomes [7]. Conventional tachycardia detection depends on cardiologists or clinical experts reading electrocardiogram (ECG) signals. Due to limited numbers of measurements and the intermittent nature of the diseases, the symptoms of tachycardia might not be captured when ECGs are recorded in hospitals [8]. Therefore, continuous monitoring enables clinicians to early diagnose, predict the disease, and have enough time to prevent patients from deteriorating.

Recently, several hospitals have attempted to utilize wearable devices for continuous monitoring of vital signs such as heart rate, respiration rate, and oxygen saturation (SpO_2) [9,10]. The adoption of wearable devices in hospitals facilitates the acquisition of patient status anywhere and anytime to reduce the workload of nurses. Compared with the use of single-threshold alarm monitoring devices and commonly used early warning scores defined by clinical experts [11], machine learning methods can automatically discover patterns and relationships within data without human instructions. Thus, machine learning has been proven as an effective clinical tool to identify abnormal events or provide early warning of diseases based on electronic health record, biomarker, gene expression, and imaging data [12-14]. Forkan et al [15] leveraged a hidden Markov model to predict 7 clinical onsets, including tachycardia onset, and further improved performance by using random forest algorithms to forecast events within 1 to 2 hours [16]. Lee et al [17] developed an artificial neural network to predict ventricular tachycardia within 1 hour. Szep et al [18] utilized an archetypal cardiac monitoring system with regression and boosting models to detect arrhythmia and predict the fatal arrhythmia several minutes before onset.

With nonlinear computation and flexible feature extraction, deep learning models show strong performances in representation learning and exploration of unknown information [19]. Researchers have recently used deep learning models for disease diagnosis and prediction based on physiological signals or electronic health records [20-22]. Since measuring and acquiring vital signs are easily measured and some open-source, labeled physiological signal (especially ECG signals) data sets are available [23,24], there exist many studies employing deep learning in cardiology [25]. Hannun et al [26] reported a convolutional neural network algorithm that detects heart arrhythmias using ECG signals acquired with a single-lead wearable sensor. Shashikumar et al [27] also presented a convolutional neural network model that detects and monitors atrial fibrillation. Teijeiro et al [28] introduced a long short-term memory (LSTM) network based on a set of features extracted from ECG records to classify normal sinus rhythm, atrial fibrillation, and anomalies. Gotlibovych et al [8] constructed a model combining a convolutional neural network and LSTM to achieve nearly real-time identification of atrial fibrillation. Cho et al [29] obtained a convolutional neural network model to predict atrial fibrillation within 4 to 6 minutes using ECG signals.

Cardiovascular diseases are complex and heterogeneous; multiple factors such as genetics, environment, age, and gender can affect the occurrence and severity of cardiovascular disease [30,31]. Age has been proven to be an independent risk factor, and being female is a greater risk factor for cardiovascular disease when elderly [31]. Few studies have attempted to develop a prediction tachycardia onset model that accounts for the patient's personal information. Respiratory dysfunction and common lung diseases, such as asthma, chronic obstructive pulmonary disease, and lung fibrosis are significantly more likely to cause cardiovascular disease [32]. Abnormal respiratory rate and its relative changes are a critical indicator to predict cardiac arrest [33], and SpO_2 has also been shown as a diagnostic marker of acute heart failure [34]. However, this useful information has not been used effectively, though it can be easily acquired with wearable sensors.

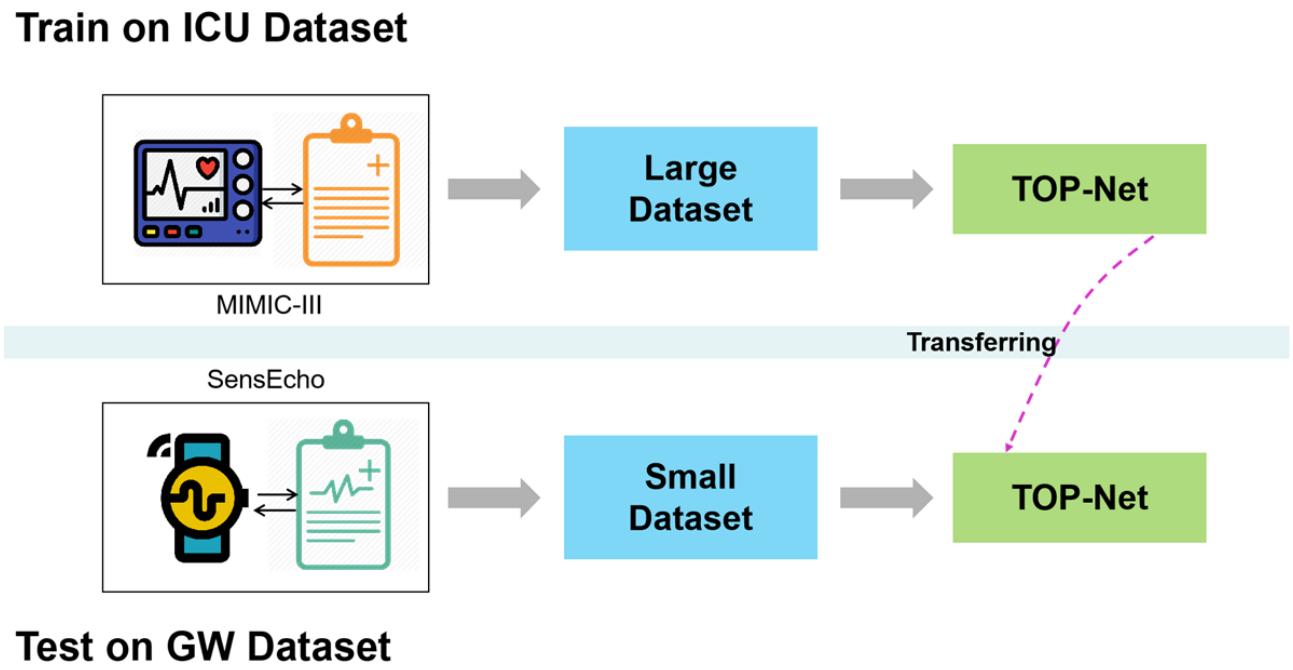
The aim of this study was to develop a bidirectional long short-term memory (BiLSTM) model—TOP-Net—that is applicable to both intensive care units and general wards [35], leverages easily accessible data, enables real-time evaluation and early prediction of tachycardia onset with a long forecast range, and is based on vital signs and electronic health record data with the following contributions: (1) combining electronic health record (sparse records) and biosensor data (high frequency records) to accomplish early prognosis and real-time prediction of tachycardia onset, and its performance of early prediction; (2) being the first to consider 2 other important vital signs and explore their different combinations being with deep learning models to predict tachycardia onset, which can improve the precision of early forecast; and (3) utilizing a large critical care data set and a model that is transferrable to real clinical scenarios wards where patients are monitored by medical-grade wearable embedded systems, for example, transferable between different countries (US to China), ethnicities (multiracial to Asian), and medical departments (intensive care unit to general ward).

Methods

Overview

We leveraged a large data set from the Medical Information Mart for Intensive Care III (MIMIC-III) [24] and its matched physiological waveform database (recorded with monitors) [36] to develop the TOP-Net model (codes available [37]). The pretrained model was transferred to a relatively small data set, from patients who were continuously monitored with a medical-grade wearable embedded system (SensEcho, Beijing SensEcho Science & Technology Co Ltd) in a real clinical environment [38]. The process is presented in Figure 1.

Figure 1. The process of developing and transferring the early tachycardia onset model, TOP-Net. GW: general ward; ICU: intensive care unit.



Methodology

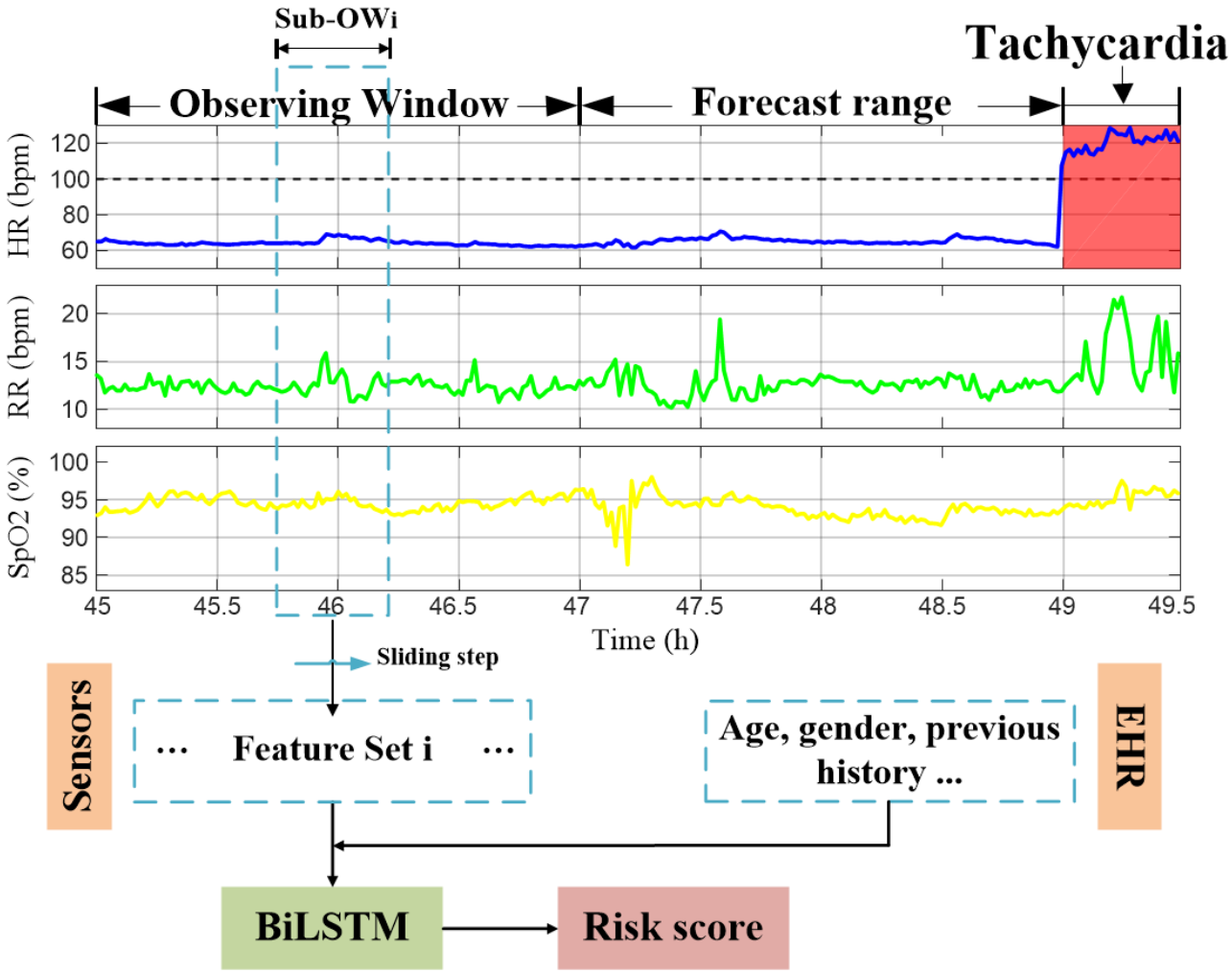
We combined 2 types of data to develop TOP-Net: (1) information from biological sensors (wearable), including heart rate, respiratory rate and SpO_2; (2) patients' personal information from electronic health records, which represents their individual health status when admitted to the hospital, including age, gender, admission type, first care unit, and history of cardiovascular disease.

TOP-Net Tachycardia Onset Early Prediction Using BiLSTM Model

Model Overview

BiLSTM [39], a sequential model, can capture the complex and multivariate dynamics in longitudinal electronic health record data and continuously collected physiological signals that is typically used in acute condition prediction, classification, and subphenotype identification [40]. We developed the model (Figure 2) using BiLSTM to take advantage of potential long-term and short-term changes and associated characteristics of physiological state.

Figure 2. An overview of TOP-Net using the cohort admission and personal measurement data in hospital. BiLSTM: bidirectional long short-term memory; EHR: electronic health record; HR: heart rate; RR: respiratory rate; SpO_2: blood oxygen saturation.



Step 1: Calculate Statistical Features

We used a BiLSTM algorithm to represent the relationship between the multiple timeseries collected by biological sensors. Data from an observing window before tachycardia onset were used to train the model. Inspired by convolutional-LSTM model [41], we designed the model to use the statistical features of the raw timeseries signals as inputs within a sliding sub–observing window. The results for all sub–observing windows were concatenated along the time and fed into the model.

We explored 8 types of statistical features—mean, standard deviance, slope, quantiles, sum, absolute energy (f_1), aggregation function of autocorrelation (f_2), and measurement of discrimination power (f_3)—that are commonly used to describe the timeseries characteristics. Herein, we focus on explaining the calculation process of f_1, f_2, and f_3.

The absolute energy of the timeseries is calculated as

The correlation of a timeseries and its time lag is described by f_2,

which is a similarity measurement index where X_i is a timeseries value at one time point, n is the length of X, σ^2 and μ are estimations of the timeseries variance and mean, respectively, and l is the time lag [42].

The nonlinearity of a timeseries is quantized using

where *lag* is a time delay operator (equal to l) [43].

Step 2: Fuse Patient Characteristics

We extracted the previously mentioned static patient information which was merged with the statistical features. The concatenated vectors were normalized and input to the BiLSTM model.

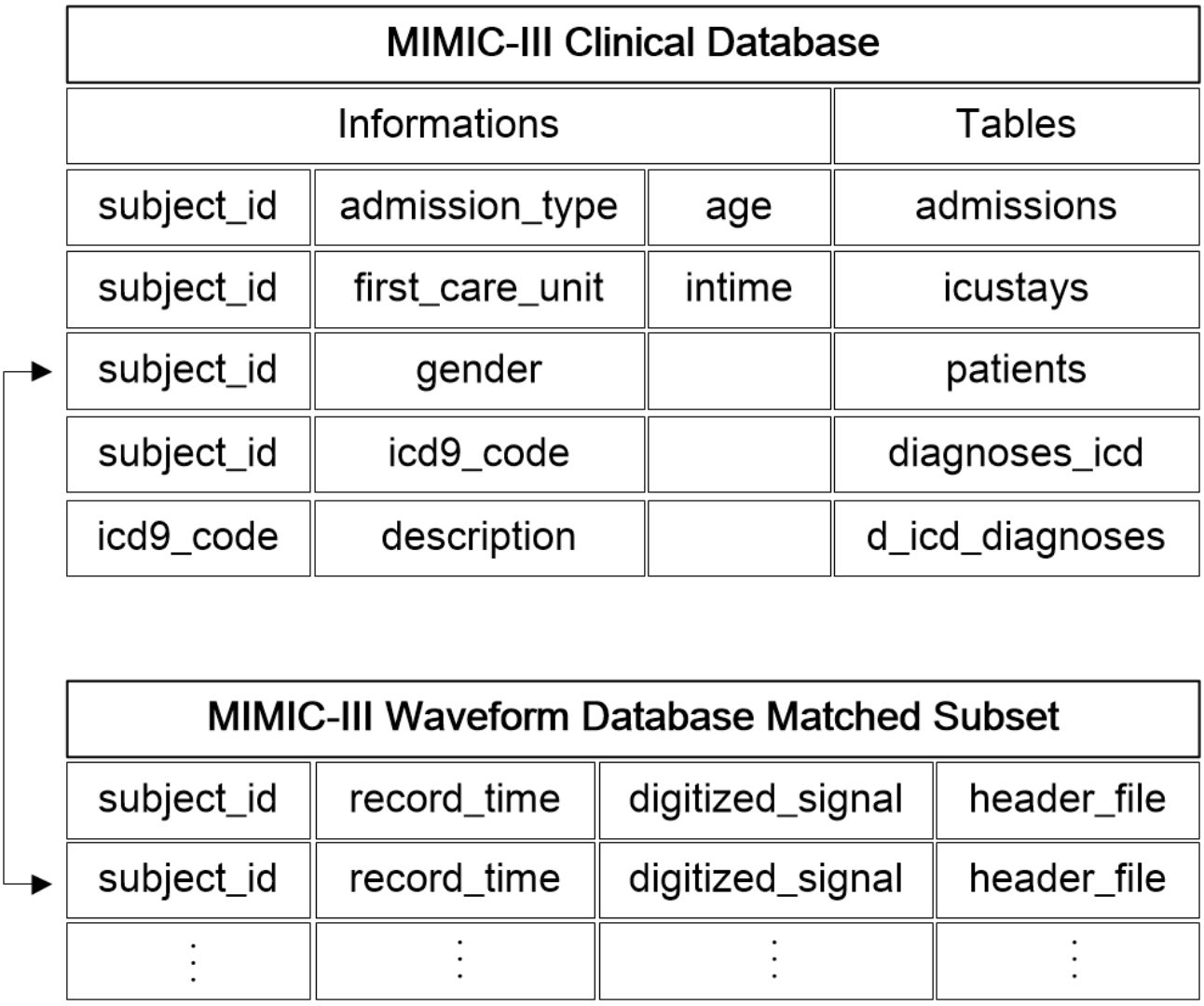
Step 3: Obtain Tachycardia Onset Risk Score

In this step, TOP-Net determines a real-time risk score that evaluates an individual risk probability of tachycardia onset. When the risk score continuously exceeds the threshold set by the doctor for a period of time, the caregiver is alerted.

Medical Information Mart for Intensive Care (MIMIC)

MIMIC III is a large, publicly available critical care database (version 1.4 [24]), with 38,557 adult patients' (52,955 ICU admissions) detailed hospital information such as demographic information, laboratory test results, and diagnosis codes. Patients' multiple physiological signals (waveforms) and corresponding numeric format of vital signs are stored in the MIMIC III Waveform Database, which contains 10,282 patients' time alignment information and 22,247 numeric records that can be matched to the clinical database [36]. The basic information is stored in the tables of *admissions*, patients' hospital admission information; *icustays*, ICU transfer (in and out) information; *patients*, individual birth and death dates; and *diagnoses_icd*, diagnosis codes during hospitalization. All of the tables can be associated with *subject_id*, a unique identity of patients. The waveform database includes the header files (name, unit, and recording frequency) and segments of recordings (numeric signals). Figure 3 presents the method used to link tables of information with the temporal waveforms.

Figure 3. The connection between clinical and waveform information in the MIMIC-III database.



Continuous Monitoring Database for the General Ward

The use of general ward data was approved by the ethics committee of the General Hospital of PLA (S2018-095-01). In the general ward, we utilized a SensEcho medical-grade monitoring system, which can monitor patients anytime and anywhere. SensEcho contains 3 parts (Figure 4): a wearable multisensor system unit, a wireless network and data transmission unit, and a central monitoring system [35,38]. The multisensors include a single-lead ECG sensor (200 Hz), a sensor for respiratory inductive plethysmography (25 Hz), a noninvasive photoplethysmogram sensor for SpO_2 monitoring (1 Hz) based on near-infrared spectroscopy, and a posture recognition sensor using a 3-axis accelerometer. These signals are collected and stored in a data logger. The logger has an ultra–low power Wi-Fi module and supports long-term data transmission by relying upon hospital networks. The central monitoring system receives information, processes data, and delivers and displays information. The algorithms deployed on the system included signal quality evaluation, signal processing, real-time abnormal event monitoring and early prediction, and patients' health assessment, which were packaged as a toolkit (Midas). The accuracy, stability, and effectiveness of our system have been validated in previous studies [44-46].

Patients admitted to the hospital were assessed by a doctor using the system. Continuous monitoring physiological signals were transmitted to the hospital server and the data in numeric format were acquired based on the waveform processing function in Midas. The clinical information was stored separately in the hospital information system. Data from the different sources were linked (Figure 5) using *patient_id*, a unique identification of patients similar to *subject_id* in MIMIC III.

Figure 4. Overview of the SensEcho system.

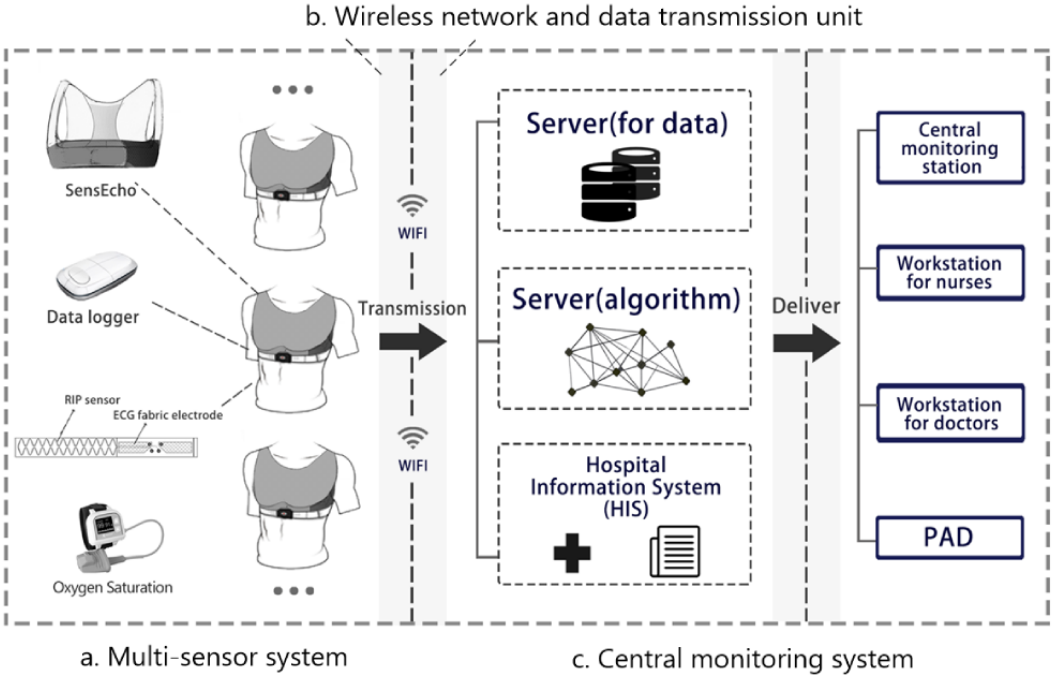


Figure 5. The connection between clinical and waveform information monitored by SensEcho.

PLAGH Hospital Information System			
Informations			Tables
patient_id	gender		pat_master_index
patient_id	admission_type	age	pat_visit
patient_id	first_care_unit	intime	transfer
patient_id	icd_code (9/10)		diagnosis
icd_code	description		d_icd_diagnosis

Wearable Continuous Monitoring Database			
patient_id	waveform	digitized_signal	header_file
patient_id	waveform	digitized_signal	header_file
⋮	⋮	⋮	⋮

Tachycardia Onset Diagnostic Criteria

Diagnostic tachycardia onset criteria were determined by 3 clinical experts from the Emergency Department, the general ward, and surgical ICU. A tachycardia event was defined as any of the following: (1) heart rate above 100 bpm sustained over 30 minutes; (2) heart rate above 130 bpm sustained over 20 minutes; (3) heart rate above 150 bpm sustained over 5 minutes. The initial timepoint meeting of any of these conditions was recognized as tachycardia onset.

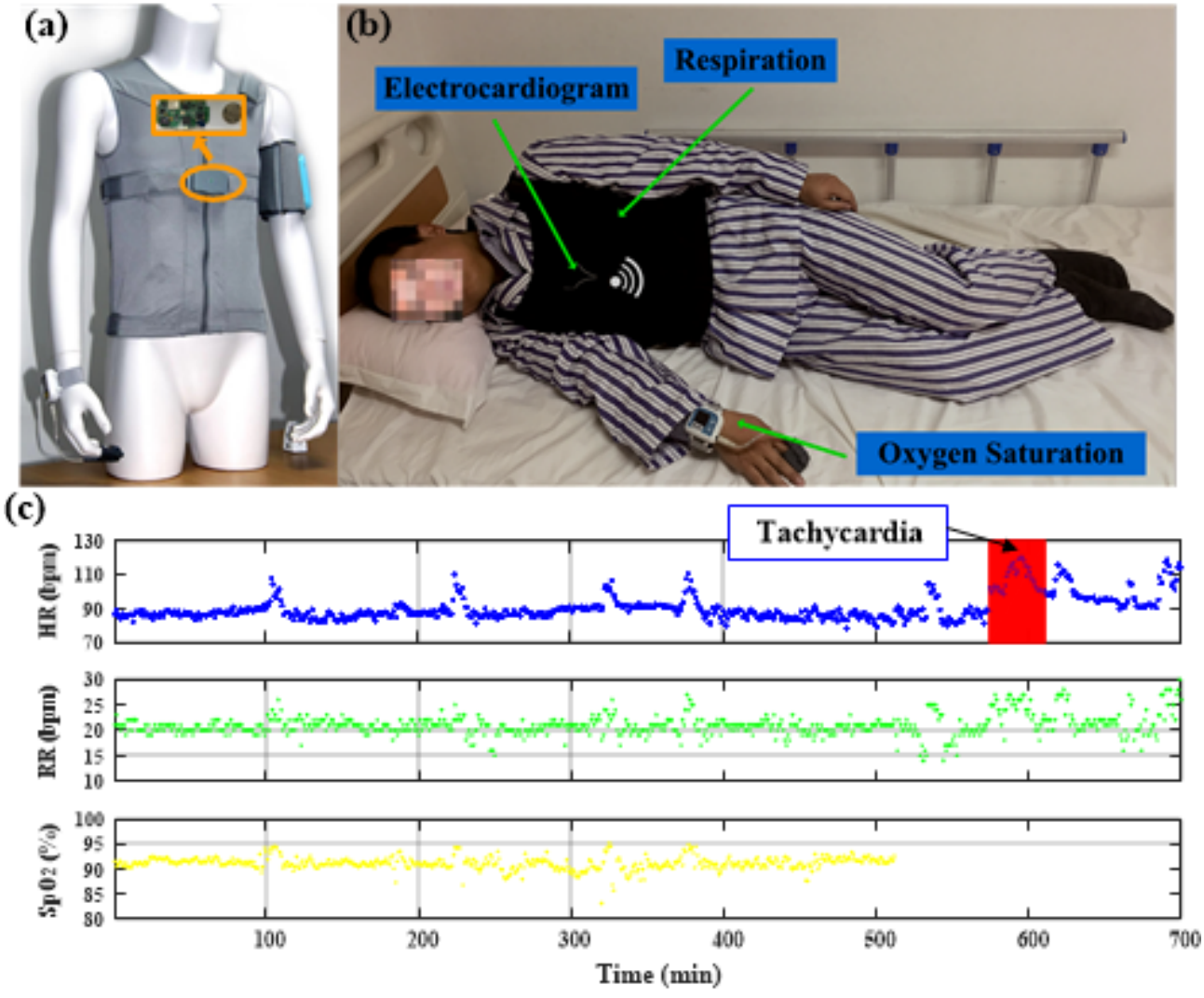
Experiments

Data Set

In the ICU environment, we selected 5699 patients with the following criteria: age over 18 years old, admitted to the hospital and ICU for the first time, monitoring data longer than 14 hours with heart rate, respiratory rate, and SpO_2 recordings. The size of the observing window was chosen as 2 hours, which was used to extract the statistical features. The negative sample set was built by extracting information in the observing window with a 1-hour sliding step throughout monitoring for patients without tachycardia. The positive sample set was acquired by selecting the same features in the observing window before the occurrence of tachycardia with a forecast range. To balance the ratio of positive and negative samples, we kept extracting positive samples with a 5-minute delay based on the former (for target replication), which is a method used in a previous study [47]. The data were downsampled from per second to per minute by averaging. If more than 30% were null or 0 values of all variables at a certain time, the missing values were filled using the forward interpolation method. We randomly picked the number of negative samples close to the positive samples to further decrease class imbalances. There were 2748 and 2130 negative and positive samples, respectively.

In the general ward, we deployed the wearable grade monitoring system (Figure 6a) in a cardiovascular disease department in January 2018. We collected data from 367 patients for research. The inclusion criteria for monitoring duration was reduced to from 14 hours to 4 hours to take into account patient length of stay. A total of 259 patients were included, and 2300 negative samples and 270 positive samples were extracted. Figure 6b shows a patient wearing a multisensor shirt, and Figure 6c shows an example of a patient encountering tachycardia.

Figure 6. Continuous monitoring using (a) SensEcho system with (b) example of a patient with sensors attached, and (c) sample data. HR: heart rate; RR: respiratory rate; SpO_2: blood oxygen saturation.



Developing the Prediction Model

In the early prediction model, developed from the MIMIC-III data set, predictions (forecast ranges) with TOP-Net were explored from 0 hour to 6 hours with a 2-hour interval. A total of 21 statistical features were included (Table 1). The size of sub–observing window and sliding step were individually set to 20 minutes and 5 minutes, respectively. We calculated all statistical values in sub–observing windows, sequentially amalgamated, and fed them into the model. The data set was randomly split to 80% of the training set and 20% of the testing set according to the patient's hospitalization number. The 5-fold cross-validation together with random search was used to tune the hyperparameters based on the training set considering the sample size [48]. The hidden size was set to 32. We tested learning rates ranging from 1^{-4} to 1^{-2} with an interval of 1^{-4} and training epochs from 5 to 100 with an interval of 10. The best hyperparameters were determined by minimizing validation loss. We retrained the model using the optimal hyperparameters on the training set, and the performance of the model was assessed on the test set.

Table 1. Statistical features constructed in this study.

Feature type and name	Feature description
Heart rate (n=10)	
hr_mean	Mean heart rate
hr_std	Heart rate SD
hr_sum	Sum of heart rate
hr_slope	Slope of heart rate
hr_abs_energy	f_1 of heart rate
hr_c2	f_3 of heart rate with *lag*=2
hr_c3	f_3 of heart rate with *lag*=3
hr_quantiles_01	10% quantile of heart rate
hr_quantiles_03	30% quantile of heart rate
hr_quantiles_07	70% quantile of heart rate
Respiratory rate (n=5)	
resp_mean	Mean respiration rate
resp_std	Respiration rate SD
resp_slope	Slope of respiration rate
resp_abs_energy	f_1 of respiration rate
resp_c3	f_3 of respiration rate with *lag*=3
SpO_2[a] (n=5)	
spo2_mean	Mean SpO_2
spo2_std	SD of SpO_2
spo2_slope	Slope of SpO_2
spo2_c3	f_3 of SpO_2 with *lag*=3
spo2_abs_energy	f_1 of SpO_2
Together (heart rate, respiratory rate, SpO_2) (n=1)	
all_autocorrelation	Mean value of f_2 using all vital signs with the default *l*=40

[a]SpO_2: blood oxygen saturation.

Comparison With Baseline Models

To further investigate the performance of TOP-Net, we designed subexperiments 1, 2, and 3 to obtain a comprehensive assessment. In subexperiment 1, the model was acquired without considering personal information and bidirection memory functions. That is, LSTM and convolutional neural network models were obtained in a total cohort without considering the personal information of patients. The structure of the LSTM was consistent with that of a BiLSTM, and the convolutional neural network model had 2 convolutional layers. In subexperiment 2, conventional machine learning methods, including extreme gradient boosting [49], multilayer perceptron, and random forest, were compared with TOP-Net with default model parameters. In subexperiment 3, different feature

combinations were examined: (1) all vital signs, (2) heart rate, (3) heart rate and respiratory rate, and (4) heart rate and SpO_2.

Performance Evaluation Metrics

Prediction performance was measured with 6 metrics: sensitivity, specificity, accuracy, F1 score, precision, and area under the receiver operating characteristic curve (AUROC).

Model Validation and Transfer to the General Ward

The performance of TOP-Net was validated using the data collected in the general ward (small data set obtained within 1 year) by the SensEcho system. A transferrable model suitable for non-ICU patients was acquired by finetuning the ICU scenario model. The model performance was also assessed with the 6 metrics using 5-fold cross-validation due to the small sample size.

Experimental Platform

We utilized PostgreSQL (version 9.6; PostgreSQL Global Development Group) to extract the clinical data. All data processing and analyses, model development, and result visualization was performed with Python (version 3.7.1) and CUDA (version 10.0).

Results

Data Sets

Table 2 shows admission information summary statistics for the study cohorts. The patients' ages were slightly higher in the ICU cohort and most of them were admitted to the hospital for emergencies. A large proportion of patients were admitted for elective reasons in the cardiovascular disease department of our hospital. Furthermore, a higher proportion of patients had a history of cardiovascular diseases in the general ward.

Table 2. Study cohorts.

	ICU[a] cohort (n=5699)	General ward cohort (n=259)
Age (years), median (IQR)	66.15 (53.97, 77.78)	61.00 (53.00, 67.50)
Gender, n (%)		
Female	3262 (57.2)	105 (40.5)
Male	2437 (42.8)	154 (59.5)
Admission type, n (%)		
Elective	979 (17.2)	227 (87.6)
Emergency	4550 (79.8)	32 (12.4)
Urgent	170 (3.0)	—[b]
First care unit, n (%)		
Coronary care	1190 (20.9)	—
Cardiac surgery recovery	1118 (19.6)	—
Medical ICU	1501 (26.3)	—
Surgical ICU	1320 (23.2)	—
Trauma/surgical ICU	570 (10.0)	—
Cardiovascular diseases, n (%)	4933 (86.6)	234 (90.3)

[a]ICU: intensive care unit.

[b]No data.

Model Performance

Evaluation Based on the ICU Cohort

We leveraged 5-fold cross-validation to select optimal hyperparameters with the training set and assessed the performance of the model on the test set. The hyperparameter values that we selected were learning rate =0.0002, epoch=20, and batch size=64. Figure 7 and Table 3 summarize the results from subexperiment 1 and subexperiment 2. The AUROC and F1 score for TOP-Net were consistently better than those of other models, with the exception of F1 score (TOP-Net's was slightly lower than that of the LSTM model for 6 hours prediction, though TOP-Net's sensitivity was slightly higher than of the LSTM at this time).

Although the 95% CI in subexperiment 1 overlaps, TOP-Net has better performance than LSTM and convolutional neural network in each prediction range above 0.5%-1%. Therefore, fusing patient personal information and bidirection memory makes the prediction model more accurate and robust. In subexperiment 2, TOP-Net was consistently superior to the other machine learning models, especially 6 hours before tachycardia onset; TOP-Net performs well (AUROC 0.796, 95% CI 0.768-0.824; sensitivity 0.753, 95% CI 0.663-0.793; specificity 0.720, 95% CI 0.645-0.758; and F1 score 0.718).

In Table 4, the results for models using heart rate (n=10), heart rate and respiratory rate (n=15), heart rate and SpO_2 (n=15), and statistical features of all vital signs (n=21) are shown. For 2- to 6-hour forecast ranges the model with all of the features

input has the best performance with highest AUROC values. The performance is slightly reduced when inputting heart rate and respiratory rate, or heart rate and SpO_2. The performance was the worst when including only heart rate statistical features. The statistical characteristics of heart rate play a dominant role in real-time diagnosis. Furthermore, we employed the extreme gradient boosting algorithm to rank the importance of 21 designed features for a forecast range of 6 hours. The top 8 features (Figure 8) were *hr_abs_energy*, *hr_quantiles_01*, *hr_c3*, *hr_c2*, *hr_quantiles_03*, *resp_c3*, *hr_mean*, and *hr_quantiles*_07. The nonlinearity features—*hr_c3* and *hr_c2* (f_3 with *lag*=3 and *lag*=2)—were ranked third and fourth, respectively. The respiratory feature *resp_c3* was ranked sixth.

Figure 7. TOP-Net performance: (a) AUROC and (b) F1 score. AUROC: area under the receiver operating characteristic curve; CNN: convolutional neural network; LSTM: long short-term memory; XGBoost: extreme gradient boosting; MLP: multilayer perceptron; RF: random forest; TO: tachycardia onset.

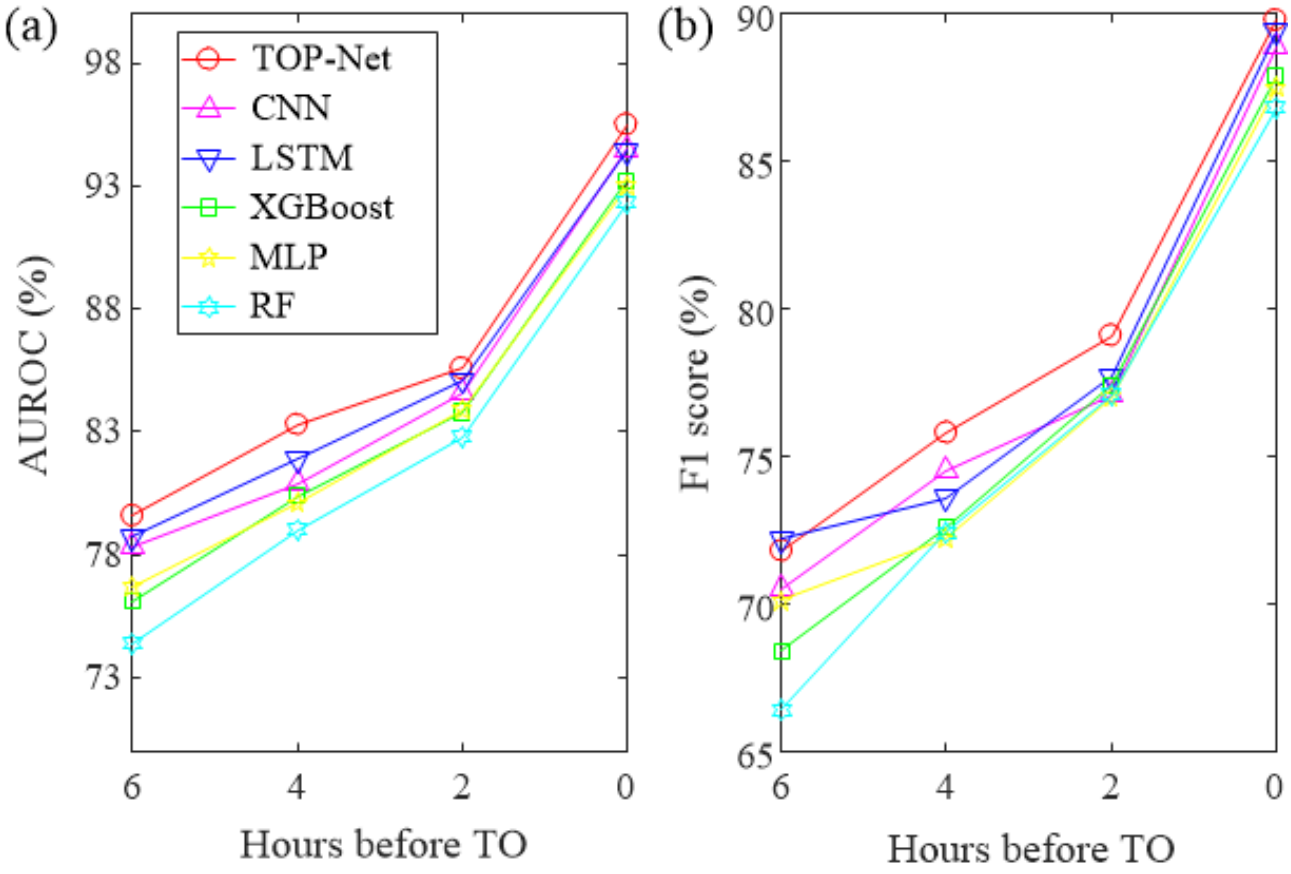


Table 3. The detailed information of performance comparison (TOP-Net vs other models).

Forecast range and model	AUROC[a] (%) (95% CI)	Accuracy (%)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	F1 score (%)	Precision (%)
0 hours						
TOP-Net	95.5 (94.2-96.8)	90.1	89.1 (81.9-91.8)	92.1 (85.9-94.3)	89.8	90.5
CNN[b]	94.5 (93.0-96.0)	89.3	85.9 (80.7-89.4)	92.3 (87.0-95.1)	88.9	92.1
LSTM[c]	94.4 (92.9-96.0)	89.8	88.9 (83.9-91.8)	90.8 (81.1-93.8)	89.4	89.9
XGBoost[d]	93.2 (91.5-94.9)	88.3	81.9 (75.7-85.6)	93.0 (87.4-96.0)	87.9	94.9
MLP[e]	93.0 (91.3-94.8)	87.9	85.6 (80.2-88.9)	89.9 (84.6-93.2)	87.5	89.5
Random forest	92.3 (90.5-94.2)	87.3	85.1 (80.2-88.8)	89.0 (82.1-92.8)	86.8	88.6
2 hours						
TOP-Net	85.6 (83.2-88.0)	79.6	77.6 (70.8-81.3)	81.6 (74.2-85.1)	79.1	80.6
CNN	84.6 (82.1-87.1)	77.6	78.6 (71.3-83.2)	77.8 (71.2-81.4)	77.1	75.6
LSTM	85.1 (82.7-87.5)	78.2	88.6 (81.0-92.0)	67.4 (56.8-71.5)	77.7	76.8
XGBoost	83.8 (81.2-86.3)	78.0	74.5 (66.7-79.1)	80.9 (73.9-84.5)	77.4	80.5
MLP	83.9 (81.4-86.4)	77.5	78.3 (71.3-82.2)	77.7 (69.9-82.0)	77.0	75.8
Random forest	82.8 (80.2-85.4)	77.7	71.5 (63.6-76.6)	82.3 (76.4-86.0)	77.1	83.7
4 hours						
TOP-Net	83.3 (80.7-85.8)	76.3	83.5 (75.5-85.9)	72.2 (63.8-74.7)	75.8	69.4
CNN	80.9 (78.2-83.7)	75.2	71.5 (63.6-76.3)	78.8 (70.0-82.5)	74.5	77.8
LSTM	81.9 (79.2-84.5)	74.2	73.1 (65.5-77.9)	76.3 (69.6-80.1)	73.6	74.1
XGBoost	80.4 (77.7-83.2)	73.4	68.1 (60.2-72.7)	78.5 (72.0-82.8)	72.6	77.8
MLP	80.1 (77.3-82.8)	72.9	73.9 (66.9-78.7)	72.0 (65.1-76.3)	72.2	70.6
Random forest	79.0 (76.1-81.9)	73.3	64.5 (60.6-71.4)	79.9 (73.4-84.8)	72.4	82.5
6 hours						
TOP-Net	79.6 (76.8-82.4)	72.1	75.3 (66.3-79.3)	72.0 (64.5-75.8)	71.8	68.6
CNN	78.3 (75.4-81.1)	70.9	79.3 (72.8-83.7)	64.1 (57.1-69.1)	70.5	63.5
LSTM	78.7 (75.9-81.5)	72.5	74.0 (67.0-78.4)	71.8 (64.1-76.0)	72.2	70.5
XGBoost	76.1 (73.1-79.0)	69.1	76.3 (69.5-75.4)	64.1 (55.1-68.9)	68.4	62.0
MLP	76.7 (73.8-79.6)	70.6	71.9 (65.1-76.7)	69.4 (61.7-74.5)	70.1	68.4
Random forest	74.4 (71.4-77.5)	67.2	69.1 (59.0-74.3)	66.7 (59.3-70.6)	66.4	63.9

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

[b]CNN: convolutional neural network.

[c]LSTM: long short-term memory.

[d]XGBoost: extreme gradient boosting.

[e]MLP: multilayer perceptron.

Table 4. Performance of TOP-Net with the different types of features.

Forecast range and feature type	AUROC[a] (%) (95% CI)	Accuracy (%)	Sensitivity (%) (95% CI)	Specificity (%) (95% CI)	F1 score (%)	Precision (%)
0 hours						
All	95.5 (94.2-96.8)	90.1	89.1 (81.9-91.8)	92.1 (85.9-94.3)	89.8	90.5
HR[b]+SpO_2[c]	95.2 (93.8-96.6)	90.4	89.4 (84.7-93.1)	91.9 (85.9-94.1)	90.1	90.8
HR+RR[d]	95.3 (93.9-96.7)	90.0	89.6 (84.7-92.3)	91.0 (84.8-94.3)	89.6	89.6
HR	95.5 (94.2-96.9)	90.1	89.1 (83.9-92.3)	92.1 (86.5-94.9)	89.8	80.6
2 hours						
All	85.6 (83.2-88.0)	79.6	77.6 (70.8-81.3)	81.6 (74.2-85.1)	79.1	80.6
HR+SpO_2	83.3 (80.8-85.9)	76.9	77.1 (70.6-81.0)	76.4 (69.4-80.4)	76.1	75.1
HR+RR	84.4 (81.9-86.9)	79.1	75.4 (69.1-79.3)	82.0 (75.3-86.5)	78.6	82.1
HR	82.9 (80.3-85.5)	76.9	78.6 (71.8-82.7)	73.9 (66.7-78.0)	76.3	74.1
4 hours						
All	83.3 (80.7-85.8)	76.3	83.5 (75.5-85.9)	72.2 (63.8-74.7)	75.8	69.4
HR+SpO_2	82.3 (79.6-84.9)	75.8	76.5 (70.0-81.3)	74.9 (67.6-79.2)	75.0	73.6
HR+RR	82.1 (79.5-84.8)	75.6	72.7 (66.4-77.7)	77.9 (70.0-82.3)	75.0	77.5
HR	80.4 (77.6-83.2)	73.6	75.5 (67.2-79.9)	72.7 (66.0-77.0)	72.9	70.5
6 hours						
All	79.6 (76.8-82.4)	72.1	75.3 (66.3-79.3)	72.0 (64.5-75.8)	71.8	68.6
HR+SpO_2	77.6 (74.7-80.5)	71.9	70.0 (62.6-74.4)	74.5 (66.5-78.6)	71.5	73.0
HR+RR	78.7 (75.8-81.5)	72.0	78.8 (71.6-83.3)	67.6 (61.2-72.0)	71.6	65.6
HR	75.5 (72.5-78.6)	70.0	67.2 (59.5-72.1)	73.3 (66.9-77.7)	69.4	71.7

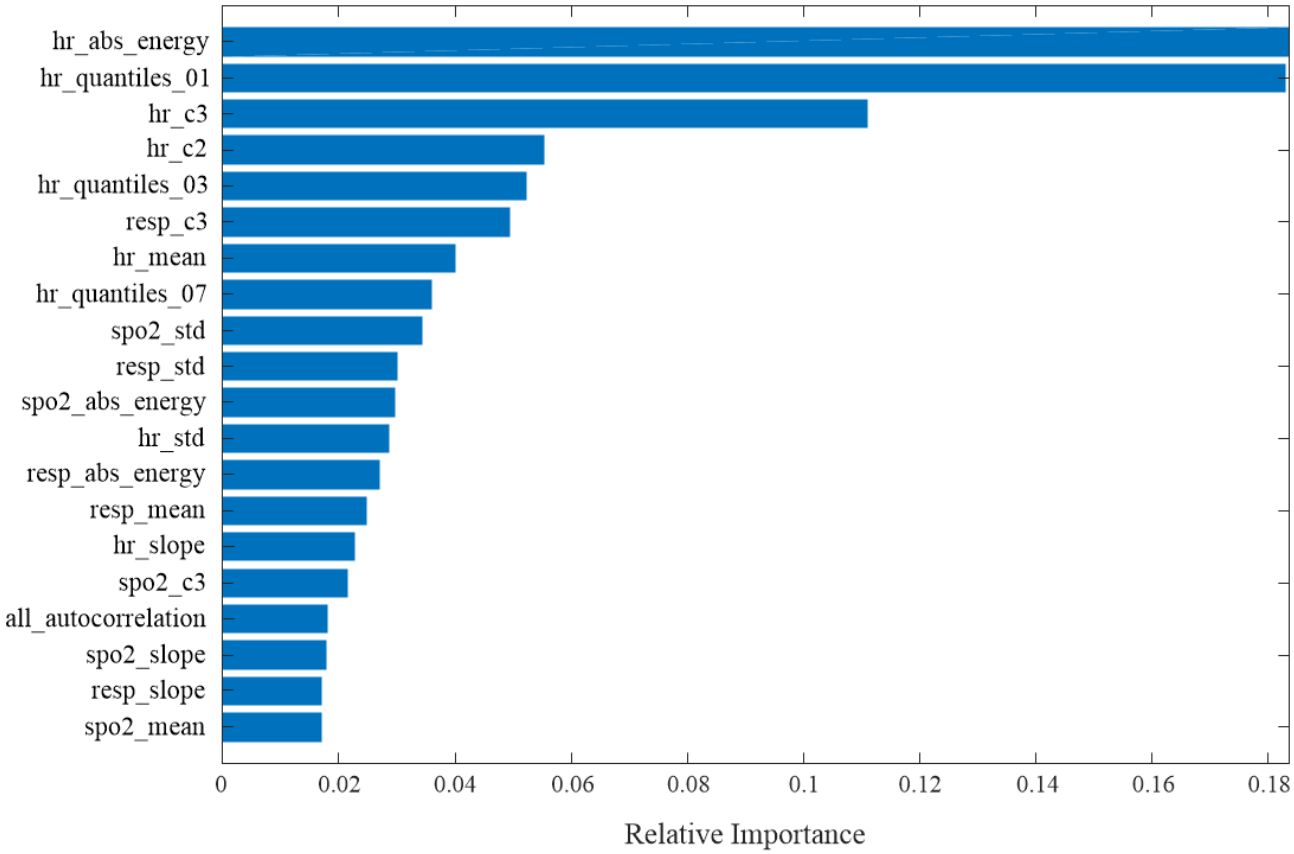
[a]AUROC: area under the receiver operating characteristic curve.

[b]HR: heart rate.

[c]SpO_2: blood oxygen saturation.

[d]RR: respiration rate.

Figure 8. Statistical feature rankings.



Model Validation in the General Ward

We assessed the performance of the model 2 hours before tachycardia onset because the interval between the tachycardia onset and the admission time to the department was short in our scenario of the general ward. Given the limited training data, we used the transfer learning method to finetune the model. The parameters were learning rate=0.0002, epoch=18, and batch size=32. The 5-fold cross-validation was also used to assess the performance and prevent possible overfitting. The retraining results can be seen in Table 5. TOP-Net had a stable outcome and outperformed the other 5 models (AUROC 0.965, accuracy 0.937, sensitivity 0.955, specificity 0.881, F1 score 0.793, and precision 0.680. Compared with the model in ICU, the difference in prediction performance might be caused by the difference in the severity of the patient's disease. Although convolutional neural network's F1 score was much higher, its sensitivity, to which clinicians pay more attention, was lower than that of TOP-Net.

Figure 9 shows real-time risk scores of tachycardia onset and an example of early tachycardia onset prediction with TOP-Net. In Figure 9a, the patient encountered a tachycardia event after admission from 675 to 725 minutes. The risk probability was assessed every 5 minutes; Figure 9b presents real-time risk. We set the alarm threshold to 0.40 with a trade-off predictive effect of sensitivity and specificity. The risk score begins to rise after the 555th minute, showing that our model can predict the tachycardia event 125 minutes beforehand.

Table 5. TOP-Net performance based on transfer learning in the general ward (2-hour forecast range).

Model	AUROC[a], mean (SD)	Accuracy (%), mean (SD)	Sensitivity (%), mean (SD)	Specificity (%), mean (SD)	F1 score (%), mean (SD)	Precision (%), mean (SD)
TOP-Net	96.5 (1.92)	93.7 (1.02)	95.5 (4.85)	88.1 (4.28)	79.3 (4.33)	68.0 (5.99)
CNN[b]	93.8 (2.02)	95.3 (1.43)	90.1 (2.88)	88.1 (8.4)	83.8 (5.38)	78.8 (9.85)
LSTM[c]	93.2 (1.89)	92.6 (0.61)	93.6 (2.76)	81.5 (5.6)	73.0 (3.4)	60.0 (4.89)
XGBoost[d]	89.9 (2.1)	92.9 (1.1)	83.4 (5.2)	82.6 (7.9)	73.7 (3.7)	66.6 (6.8)
MLP[e]	84.2 (4.1)	91.0 (0.7)	75.9 (9.6)	78.9 (9.1)	62.6 (2.0)	54.0 (2.9)
Random forest	87.3 (3.0)	92.5 (1.0)	76.6 (5.2)	86.8 (4.7)	75.0 (3.7)	73.8 (4.9)

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

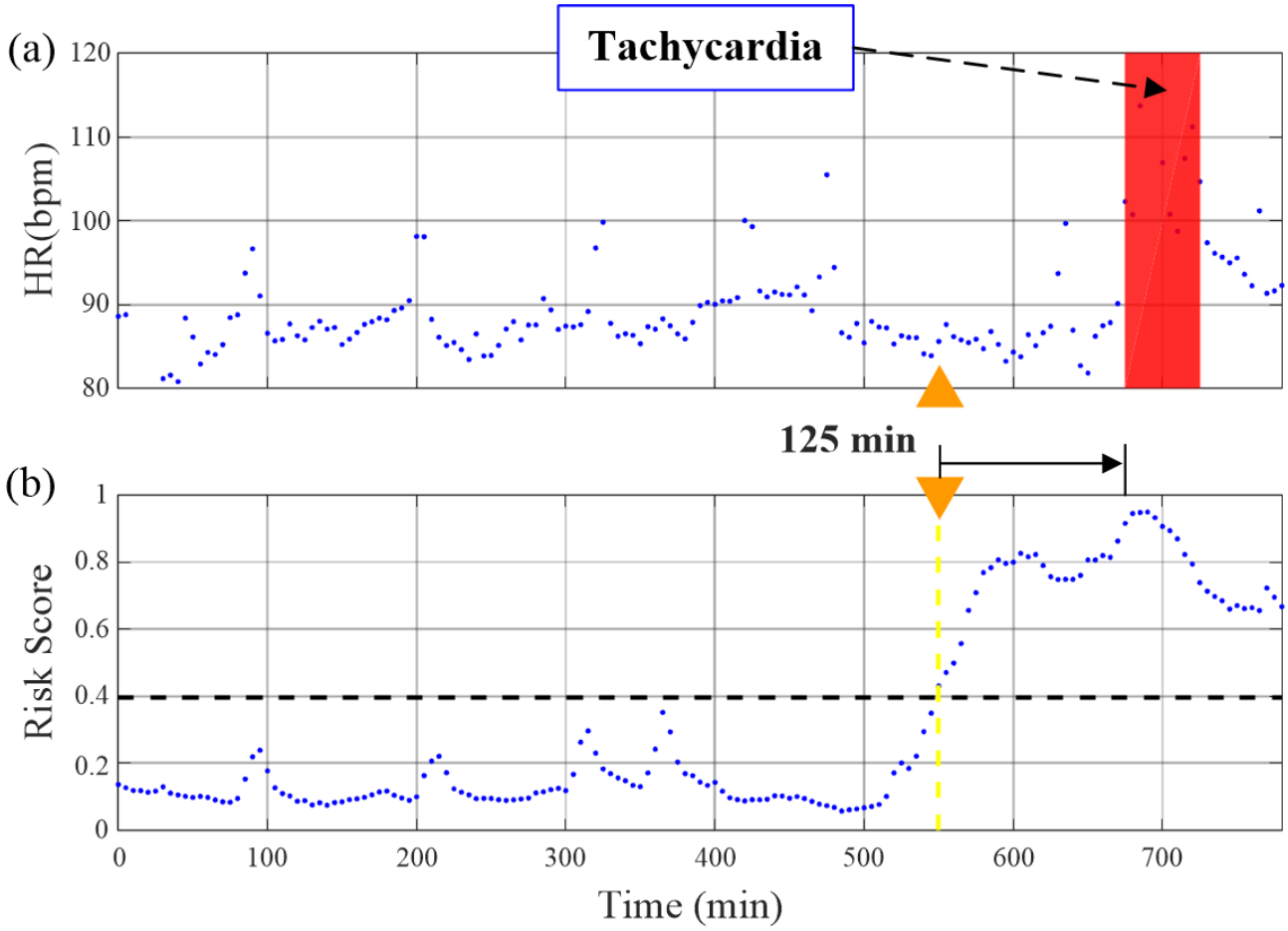
[b]CNN: convolutional neural network.

[c]LSTM: long short-term memory.

[d]XGBoost: extreme gradient boosting.

[e]MLP: multilayer perceptron.

Figure 9. Example of a tachycardia event and our risk score of predicting tachycardia onset. HR: heart rate.



Discussion

General

In this study, we developed a model using a publicly accessible data set and transferred it to a real clinical scenario. The performance of TOP-Net for predicting tachycardia onset 0 to 6 hours in advance was better than that of the baseline models (timeseries prognosis methods and conventional machine learning methods without timing characteristics); TOP-Net outperformed benchmarks of 2 deep learning models, 2 ensemble, and 1 neural network models for predictions 6 hour in advance.

Many continuous monitoring physiological status studies have indicated the deterioration of vital signs occurred more than 6 to 12 hours before serious adverse events [50]. Continuous monitoring, early prediction, and intervention tachycardia can

reduce the occurrence of heart failure, cardiac arrest, and death. This paper proposed TOP-Net, a tachycardia onset early prediction model leveraging the BiLSTM algorithm with 8 easily accessible vital signs and personal information. TOP-Net was trained using a large ICU data set and transferred to the general ward scenario with patients monitored by wearable sensors. TOP-Net has been validated to be consistently superior to the baseline models when predicting tachycardia onset from 0 to 6 hours in advance. Including patient characteristics allowed more accurate tachycardia onset prediction than those by other models without this information. Moreover, TOP-Net achieved forecasting tachycardia onset 6 hours beforehand, and the transferred model also performed well in our clinical scenario.

In recent years, some novel models for early risk prediction of adverse events have been developed based on electronic health records or physiological signals. Pan et al [51] utilized a self-correcting deep learning approach to predict whether acute kidney injury would occur in a subsequent 6 hours. Futoma et al [52] developed a multitask Gaussian process recurrent neural network classifier to early detect sepsis achieving 4 hours in advance. Tonekaboni et al [53] trained a convolutional neural network and LSTM fusion model to predict cardiac arrest from physiological signals 24 hours in advance. For tachycardia onset prediction, Lee et al [17] used an artificial neural network–based model and 104 samples to predict ventricular tachycardia 1-hour before occurrence. Yoon et al [54] adopted a random forest–based model and 1494 samples achieving detection 75 minutes in advance. Our real-time prediction model, using the deep neural architecture on 4878 sample sets, demonstrated better and more robust performance than those of multiple baseline models, which included artificial neural network and random forest models, when predicting tachycardia onset 0 to more than 6 hours beforehand.

It is necessary for clinicians to combine a patient's current symptoms, basic information, and past medical history to diagnose disease severity [55]. For example, the proportion who might have cardiovascular disease and the risk of sustained high heart rate is not the same for patients of different ages with different histories of disease. This useful information is usually recorded in electronic health records. Recently, several researchers have tried to combine the analysis of 2 kinds of materials to represent comprehensive information and improve the performance of the models: Xu et al [56] proposed a model to predict physiological decompensation and length of ICU stay by analyzing ECG and medical records data, and Nemati et al [57] employed high-resolution vital signs and electronic health records to achieve early sepsis prediction. However, little attention has been paid to tachycardia prognosis. In this paper, we integrated electronic health record and biosensor data to accomplish early prediction. The results of subexperiment 1 show that fusing electronic health record information can improve the accuracy of early prediction compared with the LSTM and convolutional neural network models.

Risk prediction is a core task in the artificial intelligence–assisted medical domain. Cardiovascular disease prediction models based on electronic health record analysis have been studied [58-60]. Doctor AI [58] requires diagnosis codes, medication codes, or procedure codes to achieve multilabel predictions including heart failure. Jin et al [60] utilized 1864 diagnostic events to train a sequential model to predict the risk of heart failure but because they were limited by the need to obtain more information, the model cannot be used in hospitals with low information integration or in homes. Deep learning models using ECG signals have also been used for predictive health care tasks [61]. While ECG signals are susceptible to interference from physical artifacts, sensors can obtain heart rate using photoplethysmography instead of ECG signals. Therefore, models based on core vital signs can easily be used and to improve prediction performance. We selected 3 vital signs and 5 types of personal information that can easily be acquired from wearable sensors and hospital information systems, respectively. TOP-Net was developed using a large data set and transferred to our actual demand scenario. The results show that it has the potential to be used in ICU and the general ward, which also can be extended to home use. Table 6 presents a comparison between TOP-Net and other state-of-the-art approaches based on input information, model types, scenario for evaluating the model, sample sizes, and performance.

Table 6. Review of the performance of related algorithms.

Reference	Information	Model types	Scenario	Sample sizes	Performance
Lee et al 2016 [17]	High-frequency vital signs (1)	Nontemporal, classic machine learning	ICU[a]	52 (positive records); 52 (negative records)	1 hour before ventricular tachycardia: sensitivity 88%; specificity 82%; AUROC[b] 93%
Forkan et al 2017 [16]	High-frequency vital signs (6)	Nontemporal, classic machine learning	ICU	4893 (positive and negative records)	1-2 hours before tachycardia onset: accuracy 95.85%
Yoon et al 2019 [54]	High-frequency vital signs (3)	Nontemporal, classic machine learning	ICU	787 (positive records); 707 (negative records)	75 minutes before tachycardia onset: accuracy 84.7%-78.2%; AUROC 92.1%-84.2 %
TOP-Net	High-frequency vital signs (3) and electronic health record data (5)	Temporal, deep learning	ICU and the general ward	2130+270 (positive records); 2748+2300 (negative records)	6 hours before tachycardia onset: accuracy 72.1%; AUROC 79.6%

[a]ICU: intensive care unit.

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

Limitations

This study had some limitations. Because SensEcho was deployed in the clinic for only 1 year after our research project began, the limited data collected prevented us from directly developing a general ward model. Moreover, interventions such as beta-blocker medication may affect the occurrence of tachycardia onset and cause it to not be captured by the input features. Electronic health records contain rich information such as laboratory tests, clinical orders, and nursing notes that can characterize a patient's health status and depict the trajectory of diseases. Further studies involving the integration of multivariate timeseries from electronic health records are expected to improve the prediction performance of tachycardia onset, and more data from the general ward for TOP-Net performance evaluation are required.

Conclusions

TOP-Net for real-time evaluation and early prediction of the risk of tachycardia onset, which made it possible to achieve an early forecast of tachycardia onset 6 hours in advance with clinically acceptable performance. TOP-Net was assessed using 6 metrics, 3 subexperiments, different prediction times from 0 to 6 hours. The comparison between the TOP-Net and the other 5 approaches (2 deep learning models, 2 ensemble models, and 1 artificial neural network model) showed that TOP-Net was superior to the other models. The model with personal information from electronic health records had better performance than those without. The easily accessible input data of the model (3 vital signs and 5 types of personal information) and the good performance of the transferred model in the general ward indicated the early prediction of tachycardia onset using wearable sensors is possible in hospitals or houses.

Acknowledgments

We thank Yunkai Yu (Beijing Institute of Technology) for helping us train models and arrange data. We also thank Dr Alistair Johnson and Dr Tom Pollard (Massachusetts Institute of Technology) for useful suggestions and comments. This project was funded by National Key Research and Development Program (2016YFC1304305), National Natural Science Foundation of China (61471398), Beijing Municipal Science and Technology Project (Z181100001918023), Big Data Research and Development Project of Chinese PLA general hospital (2018MBD-009, 2018MBD-058) and in part by the China Education and Research Network Innovation Project (NGII20160701).

Authors' Contributions

This work was performed during ZY's internship at Beijing SensEcho Science & Technology Co Ltd as a PhD candidate at University of California. All authors came up with the study concept. XL, TL, and ZZ contributed to collecting data, designing model**s,** and drafting the manuscript. P-CK, HX, and PL contributed to further analyzing and interpreting data. P-CK, ZY, KL, and YLN contributed to cleaning data and revising the manuscript. WY and DL contributed to statistical analysis. All coauthors had the opportunity to comment on the manuscript before submission and approved the final version for submission.

Conflicts of Interest

None declared.

References

1. Awtry E, Jeon C, Ware MG. Blueprints Cardiology. United States: Lippincott Williams & Wilkins; 2006:1-20.
2. Arzbaecher R, Bump T, Jenkins J, Glick K, Munkenbeck F, Brown J, et al. Automatic tachycardia recognition. Pacing Clin Electrophysiol 1984 May;7(3 Pt 2):541-547. [doi: 10.1111/j.1540-8159.1984.tb04948.x] [Medline: 6204312]
3. Au-Yeung WM, Reinhall PG, Bardy GH, Brunton SL. Development and validation of warning system of ventricular tachyarrhythmia in patients with heart failure with heart rate variability data. PLoS One 2018 Nov 14;13(11):e0207215 [FREE Full text] [doi: 10.1371/journal.pone.0207215] [Medline: 30427880]
4. Srinivasan NT, Schilling RJ. Sudden cardiac death and arrhythmias. Arrhythm Electrophysiol Rev 2018;7(2):111. [doi: 10.15420/aer.2018:15:2]
5. Wang TJ, Larson MG, Levy D, Vasan RS, Leip EP, Wolf PA, et al. Temporal relations of atrial fibrillation and congestive heart failure and their joint influence on mortality. Circulation 2003 Jun 17;107(23):2920-2925. [doi: 10.1161/01.cir.0000072767.89944.6e]
6. Miyasaka Y, Barnes ME, Bailey KR, Cha SS, Gersh BJ, Seward JB, et al. Mortality trends in patients diagnosed with first atrial fibrillation: a 21-year community-based study. J Am Coll Cardiol 2007 Mar 06;49(9):986-992 [FREE Full text] [doi: 10.1016/j.jacc.2006.10.062] [Medline: 17336723]
7. Gardner-Thorpe J, Love N, Wrightson J, Walsh S, Keeling N. The value of Modified Early Warning Score (MEWS) in surgical in-patients: a prospective observational study. Ann R Coll Surg Engl 2006 Oct;88(6):571-575. [doi: 10.1308/003588406x130615]
8. Gotlibovych I, Crawford S, Goyal D. End-to-end deep learning from raw sensor data: atrial fibrillation detection using wearables. arXiv. Preprint posted online on July 27, 2018 [FREE Full text]

9. Dunn J, Runge R, Snyder M. Wearables and the medical revolution. Per Med 2018 Sep;15(5):429-448 [FREE Full text] [doi: 10.2217/pme-2018-0044] [Medline: 30259801]
10. Kroll RR, McKenzie ED, Boyd JG, Sheth P, Howes D, Wood M, WEARable Information Technology for hospital INpatients (WEARIT-IN) study group. Use of wearable devices for post-discharge monitoring of ICU patients: a feasibility study. J Intensive Care 2017 Nov 21;5(1):64-68 [FREE Full text] [doi: 10.1186/s40560-017-0261-9] [Medline: 29201377]
11. Alam N, Hobbelink E, van Tienhoven A, van de Ven P, Jansma E, Nanayakkara P. The impact of the use of the Early Warning Score (EWS) on patient outcomes: a systematic review. Resuscitation 2014 May;85(5):587-594. [doi: 10.1016/j.resuscitation.2014.01.013] [Medline: 24467882]
12. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 05;7(299):299ra122-299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
13. 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]
14. Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. N Engl J Med 2017 Jun 29;376(26):2507-2509. [doi: 10.1056/nejmp1702071]
15. Forkan ARM, Khalil I. A probabilistic model for early prediction of abnormal clinical events using vital sign correlations in home-based monitoring. 2016 Presented at: 2016 IEEE International Conference on Pervasive Computing and Communications; March 14-19; Sydney, Australia p. 1-9. [doi: 10.1109/percom.2016.7456519]
16. Forkan ARM, Khalil I, Atiquzzaman M. ViSiBiD: a learning model for early discovery and real-time prediction of severe clinical events using vital signs as big data. Computer Networks 2017 Feb;113:244-257. [doi: 10.1016/j.comnet.2016.12.019]
17. Lee H, Shin S, Seo M, Nam G, Joo S. Prediction of ventricular tachycardia one hour before occurrence using artificial neural networks. Sci Rep 2016 Aug 26;6(1):32390-32397 [FREE Full text] [doi: 10.1038/srep32390] [Medline: 27561321]
18. Szep J, Hariri S, Khalpey Z. Predictive diagnosis of fatal heart rhythms using wearables. 2019 Presented at: 2019 Spring Simulation Conference; April 29-May 2; Tucson, Arizona p. 1-10. [doi: 10.23919/springsim.2019.8732885]
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
20. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019 Jan 7;25(1):65-69 [FREE Full text] [doi: 10.1038/s41591-018-0268-3] [Medline: 30617320]
21. 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. [doi: 10.1109/jbhi.2017.2767063]
22. Ghassemi M, Pimentel MAF, Naumann T. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. 2015 Presented at: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; January 25-30; Austin, Texas.
23. Moody G, Mark R. The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 2001;20(3):45-50. [doi: 10.1109/51.932724] [Medline: 11446209]
24. 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):160035-160039 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
25. Bizopoulos P, Koutsouris D. Deep learning in cardiology. IEEE Rev Biomed Eng 2019;12:168-193. [doi: 10.1109/rbme.2018.2885714]
26. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019 Jan;25(1):65-69 [FREE Full text] [doi: 10.1038/s41591-018-0268-3] [Medline: 30617320]
27. Shashikumar SP, Shah AJ, Li Q. A deep learning approach to monitoring and detecting atrial fibrillation using wearable technology. 2017 Presented at: IEEE EMBS International Conference on Biomedical & Health Informatics (BHI); February 16-19; Orlando, Florida p. 141-144. [doi: 10.1109/bhi.2017.7897225]
28. Teijeiro T, García CA, Castro D. Arrhythmia classification from the abductive interpretation of short single-lead ECG records. 2017 Presented at: Computing in Cardiology; September 24-27; Rennes, France p. 1-4. [doi: 10.22489/cinc.2017.166-054]
29. Cho J, Kim Y, Lee M. Prediction to atrial fibrillation using deep convolutional neural networks. In: Rekik I, Unal G, Adeli E, Park S, editors. Predictive Intelligence in Medicine. Cham: Springer; Sep 16, 2018:164-171.
30. Krittanawong C, Zhang H, Wang Z, Aydar M, Kitai T. Artificial intelligence in precision cardiovascular medicine. J Am Coll Cardiol 2017 May 30;69(21):2657-2664 [FREE Full text] [doi: 10.1016/j.jacc.2017.03.571] [Medline: 28545640]
31. Rodgers JL, Jones J, Bolleddu SI, Vanthenapalli S, Rodgers LE, Shah K, et al. Cardiovascular risks associated with gender and aging. J Cardiovasc Dev Dis 2019 Apr 27;6(2):19 [FREE Full text] [doi: 10.3390/jcdd6020019] [Medline: 31035613]
32. Carter P, Lagan J, Fortune C, Bhatt DL, Vestbo J, Niven R, et al. Association of cardiovascular disease with respiratory disease. J Am Coll Cardiol 2019 May 07;73(17):2166-2177 [FREE Full text] [doi: 10.1016/j.jacc.2018.11.063] [Medline: 30846341]

33. Cretikos MA, Bellomo R, Hillman K, Chen J, Finfer S, Flabouris A. Respiratory rate: the neglected vital sign. Med J Aust 2008 Jun 02;188(11):657-659. [doi: 10.5694/j.1326-5377.2008.tb01825.x] [Medline: 18513176]
34. Masip J, Gayà M, Páez J, Betbesé A, Vecilla F, Manresa R, et al. Pulse oximetry in the diagnosis of acute heart failure. Revista Española de Cardiología (English Edition) 2012 Oct;65(10):879-884. [doi: 10.1016/j.rec.2012.02.021]
35. Lan K, Liu X, Xu H. DeePTOP: personalized tachycardia onset prediction using bi-directional LSTM in wearable embedded systems. 2019 Presented at: International Conference on Embedded Wireless Systems and Networks; February 25-27; Beijing, China p. 216-217.
36. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 2000 Jun 13;101(23):E215-E220. [doi: 10.1161/01.cir.101.23.e215] [Medline: 10851218]
37. TOP-Net. GitHub. URL: https://github.com/liuxiaoliXRZS/TOP-Net [accessed 2021-04-09]
38. Xu H, Li P, Yang Z, Liu X, Wang Z, Yan W, et al. Construction and application of a medical-grade wireless monitoring system for physiological signals at general wards. J Med Syst 2020 Sep 04;44(10):182-115 [FREE Full text] [doi: 10.1007/s10916-020-01653-z] [Medline: 32885290]
39. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
40. Xiao, 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]
41. Shi X, Chen Z, Wang H, Yeung DY, Wong WK, Woo WC. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. 2015 Presented at: Advances in neural information processing systems; December 7-12; Montreal, Canada p. 802-810 URL: https://papers.nips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf
42. Bence JR. Analysis of short time series: correcting for autocorrelation. Ecology 1995;76(2):628-639. [doi: 10.2307/1941218]
43. Schreiber T, Schmitz A. Discrimination power of measures for nonlinearity in a time series. Phys Rev E 1997 May 1;55(5):5443-5447. [doi: 10.1103/physreve.55.5443]
44. Li P, Yang Z, Yan W, Yan M, He M, Yuan Q, et al. Mobicardio: a clinical-grade mobile health system for cardiovascular disease management. 2019 Presented at: IEEE International Conference on Healthcare Informatics; June 10-13; Xi'an, China p. 1-6. [doi: 10.1109/ichi.2019.8904641]
45. Zhang Y, Yang Z, Zhang Z, Liu X, Cao D, Li P, et al. Automated sleep period estimation in wearable multi-sensor systems. In: Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems. 2018 Presented at: ACM SenSys 2018; November 4-7; Shenzhen, China p. 305-306. [doi: 10.1145/3274783.3275155]
46. Zhang Y, Yang Z, Lan K, Liu X, Zhang Z, Li P, et al. Sleep stage classification using bidirectional lstm in wearable multi-sensor systems. 2019 Presented at: IEEE Conference on Computer Communications Workshops; April 29-May 2; Paris. [doi: 10.1109/infcomw.2019.8845115]
47. Futoma J, Hariharan S, Heller K, Sendak M, Brajer N, Clement M, et al. An improved multi-output gaussian process RNN with real-time validation for early sepsis detection. In: Proceedings of the 2nd Machine Learning for Healthcare Conference. 2017 Presented at: Machine Learning for Healthcare Conference; August 18-19; Boston, Massachusetts p. 243-254 URL: http://proceedings.mlr.press/v68/futoma17a/futoma17a.pdf
48. Ge W, Huh JW, Park YR, Lee JH, Kim YH, Turchin A. An interpretable ICU mortality prediction model based on logistic regression and recurrent neural networks with LSTM units. AMIA Annu Symp Proc 2018;2018:460-469 [FREE Full text] [Medline: 30815086]
49. Chen T, Carlos G. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2016. 2016 Presented at: International Conference on Knowledge Discovery and Data Mining; August 13-17; San Francisco, California p. 785-794. [doi: 10.1145/2939672.2939785]
50. Welch J, Kanter B, Skora B, McCombie S, Henry I, McCombie D, et al. Multi-parameter vital sign database to assist in alarm optimization for general care units. J Clin Monit Comput 2016 Dec 6;30(6):895-900 [FREE Full text] [doi: 10.1007/s10877-015-9790-8] [Medline: 26439830]
51. Pan Z, Du H, Ngiam KY, Wang F, Shum P, Feng M. A self-correcting deep learning approach to predict acute conditions in critical care. arXiv. Preprint posted online on July 14, 2019 [FREE Full text]
52. Futoma J, Harihan S, Heller K. Learning to detect sepsis with a multitask Gaussian process RNN classifier. 2017 Presented at: 34th International Conference on Machine Learning; August 6-11; Sydney, Australia.
53. Tonekaboni S, Mazwi M, Laussen P, Eytan D, Greer R, Goodfellow SD, et al. Prediction of cardiac arrest from physiological signals in the pediatric ICU. In: Proceedings of the 3rd Machine Learning for Healthcare Conference. 2018 Presented at: 3rd Machine Learning for Healthcare; August 16-18; Stanford, California p. 534-550.
54. Yoon JH, Mu L, Chen L, Dubrawski A, Hravnak M, Pinsky MR, et al. Predicting tachycardia as a surrogate for instability in the intensive care unit. J Clin Monit Comput 2019 Dec 14;33(6):973-985 [FREE Full text] [doi: 10.1007/s10877-019-00277-0] [Medline: 30767136]
55. Ma F, Gao J, Suo Q, You Q, Zhou J, Zhang A. Risk prediction on electronic health records with prior medical knowledge. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018

Presented at: International Conference on Knowledge Discovery & Data Mining; August 19-20; London, United Kingdom p. 1910-1919. [doi: 10.1145/3219819.3220020]
56. Xu Y, Biswal S, Deshpande SR, Maher KO, Sun J. RAIM: Recurrent attentive and intensive model of multimodal patient monitoring data. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018 Presented at: International Conference on Knowledge Discovery & Data Mining; August 19-20; London, United Kingdom p. 2565-2573. [doi: 10.1145/3219819.3220051]
57. Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Crit Care Med 2018 Apr;46(4):547-553 [FREE Full text] [doi: 10.1097/CCM.0000000000002936] [Medline: 29286945]
58. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. Doctor AI: predicting clinical events via recurrent neural networks. JMLR Workshop Conf Proc 2016 Aug;56:301-318 [FREE Full text] [Medline: 28286600]
59. 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]
60. Jin B, Che C, Liu Z, Zhang S, Yin X, Wei X. Predicting the risk of heart failure with EHR sequential data modeling. IEEE Access 2018;6:9256-9261. [doi: 10.1109/access.2017.2789324]
61. Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and challenges of deep learning methods for electrocardiogram data: a systematic review. Comput Biol Med 2020 Jul;122:103801. [doi: 10.1016/j.compbiomed.2020.103801] [Medline: 32658725]

Abbreviations

AUROC: area under the receiver operating characteristic curve
BiLSTM: bidirectional long short-term memory
ECG: electrocardiogram
ICU: intensive care unit
LSTM: long short-term memory
MIMIC-III: Medical Information Mart for Intensive Care III
SpO_2: blood oxygen saturation

Edited by G Eysenbach; submitted 19.03.20; peer-reviewed by M Feng, L Falissard, S Purkayastha; comments to author 22.04.20; revised version received 06.09.20; accepted 21.02.21; published 15.04.21.

Please cite as:
Liu X, Liu T, Zhang Z, Kuo PC, Xu H, Yang Z, Lan K, Li P, Ouyang Z, Ng YL, Yan W, Li D
TOP-Net Prediction Model Using Bidirectional Long Short-term Memory and Medical-Grade Wearable Multisensor System for Tachycardia Onset: Algorithm Development Study
JMIR Med Inform 2021;9(4):e18803
URL: https://medinform.jmir.org/2021/4/e18803
doi:10.2196/18803
PMID:33856350

Original Paper

Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study

Kecheng Zhan[1], MA; Weihua Peng[2], PhD; Ying Xiong[1], PhD; Huhao Fu[1], MA; Qingcai Chen[1,3], PhD; Xiaolong Wang[1], PhD; Buzhou Tang[1,3], PhD

[1]Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology, Shenzhen, China

[2]Baidu International Technology (Shenzhen) Co, Ltd, Shenzhen, China

[3]Peng Cheng Laboratory, Shenzhen, China

Corresponding Author:
Buzhou Tang, PhD
Key Laboratory of Network Oriented Intelligent Computation
Harbin Institute of Technology
University Town
Shenzhen, 518055
China
Phone: 86 13725525983
Email: tangbuzhou@hit.edu.cn

Abstract

Background: Family history information, including information on family members, side of the family of family members, living status of family members, and observations of family members, plays an important role in disease diagnosis and treatment. Family member information extraction aims to extract family history information from semistructured/unstructured text in electronic health records (EHRs), which is a challenging task regarding named entity recognition (NER) and relation extraction (RE), where named entities refer to family members, living status, and observations, and relations refer to relations between family members and living status, and relations between family members and observations.

Objective: This study aimed to introduce the system we developed for the 2019 n2c2/OHNLP track on family history extraction, which can jointly extract entities and relations about family history information from clinical text.

Methods: We proposed a novel graph-based model with biaffine attention for family history extraction from clinical text. In this model, we first designed a graph to represent family history information, that is, representing NER and RE regarding family history in a unified way, and then introduced a biaffine attention mechanism to extract family history information in clinical text. Convolution neural network (CNN)-Bidirectional Long Short Term Memory network (BiLSTM) and Bidirectional Encoder Representation from Transformers (BERT) were used to encode the input sentence, and a biaffine classifier was used to extract family history information. In addition, we developed a postprocessing module to adjust the results. A system based on the proposed method was developed for the 2019 n2c2/OHNLP shared task track on family history information extraction.

Results: Our system ranked first in the challenge, and the F1 scores of the best system on the NER subtask and RE subtask were 0.8745 and 0.6810, respectively. After the challenge, we further fine tuned the parameters and improved the F1 scores of the two subtasks to 0.8823 and 0.7048, respectively.

Conclusions: The experimental results showed that the system based on the proposed method can extract family history information from clinical text effectively.

(JMIR Med Inform 2021;9(4):e23587) doi:10.2196/23587

KEYWORDS

family history information; named entity recognition; relation extraction; deep biaffine attention

Introduction

Family history information plays an important role in the diagnosis and treatment of diseases, especially genetic disorders. Family history information is always embedded in electronic health records (EHRs) in a semistructured/unstructured format, which needs to be unlocked by natural language processing (NLP) technology.

In order to promote research on family history information extraction, Harvard Medical School and Mayo Clinic organized national NLP challenges on family history information extraction in 2018 and 2019. The family history information extraction task includes the following two subtasks: (1) recognizing family members, living status, and observations and (2) determining which family members the recognized living status and observations belong to, which correspond to two fundamental NLP tasks, namely named entity recognition (NER) and relation extraction (RE). The NER task is usually regarded as a sequence labeling task, while the RE task is the subsequent classification task, and they are tackled by pipeline methods.

For the NER task, traditional machine learning methods, such as hidden Markov model (HMM), conditional random field (CRF) [1], and structured support vector machine (SSVM) [2], and deep learning methods, such as Bidirectional Long Short Term Memory network (BiLSTM) CRF [3] and its variants [4,5], have been widely applied. For the RE task, the typical machine learning methods include support vector machine (SVM) [6], convolutional neural network (CNN) [7], and recurrent neural network [8]. The methods mentioned above have also been applied for clinical entity recognition and RE, such as the NLP challenges organized by i2b2 in 2009 [9], 2010 [10], 2012 [11], and 2014 [12], the NLP challenges organized by SemEval in 2015 [13] and 2016 [14], the NLP challenges organized by ShARe/CLEF in 2013 [15] and 2014 [16], and the NLP challenges organized by BioCreative/OHNLP in 2018 [17]. Most of these methods process NER and RE tasks in a pipeline way, which can suffer from error propagation [18].

A number of joint learning methods have been proposed [18,19] for NER and RE subtasks to avoid error propagation from NER to RE. In the case of family history information extraction, Shi et al [17] developed deep joint learning based on the BiLSTM that won the 2018 BioCreative/OHNLP challenge [20]. Joint learning methods generally used pretrained neural language models. Neural language models pretrained on large-scale unlabeled text have recently been proven to be surprisingly effective in many downstream tasks, and Bidirectional Encoder Representation from Transformers (BERT) [21] is one of the most popular neural language models.

In this study, we proposed a novel graph-based model with biaffine attention. Inspired by the dependency parsing task [22,23], we designed a novel graph-based schema to represent family history information and introduced deep biaffine attention [22,23] to extract family history information from clinical text. A system based on the proposed method was developed for the 2019 n2c2/OHNLP challenge on family history information extraction, and it achieved the highest F1 scores of 0.8823 on subtask1 and 0.7048 on subtask2.

Methods

Task Description

There were two subtasks in the 2019 n2c2/OHNLP challenge on family history information extraction. For subtask1, we need to recognize family members with the side of the family, living status mentioned in clinical text, and observations in the family history. All family members can be normalized to standard forms in Table 1. The property of family members named “side of family” includes the following three possible values: NA (“not applicable”), maternal, and paternal. Following the work of Shi et al [17], we compared two different strategies. The first strategy recognized three types of entities (family member, observation, and living status) and determined the “side of family” property for each family member entity through a postprocessing module. The second strategy recognized five types of entities (NA, maternal, paternal, observation, and living status), directly determining the “side of family” property of family members.

For subtask2, we need to extract the relations between family members, observations, and living status. Living status is used to represent the health status of family members, and it has the two properties of “Alive” and “Healthy.” Each property was measured by a real-valued score (yes: 2, NA: 1, and no: 0). The total living status score of family members was their alive score multiplied by their health score. We also need to predict the negation information (Negated and Non_Negated) for each observation, that is, to judge whether the family members have certain diseases or not.

Table 1. Normalized family member names.

Degree	Normalized family member names
1	Father, Mother, Parent, Sister, Brother, Daughter, Son, and Child
2	Grandmother, Grandfather, Grandparent, Cousin, Sibling, Aunt, and Uncle

Data Statistics

We conducted experiments on the corpus provided by the 2018 and 2019 n2c2/OHNLP shared task tracks on family history information extraction. The training set of the 2019 n2c2/OHNLP shared task together with the test set of the 2018 BioCreative/OHNLP shared task was used as the final training set of 149 EHRs for model training. The test data set of the 2019 n2c2/OHNLP shared task, including 117 EHRs, was used for the model test. During model training, we randomly selected a development set of 14 EHRs from the training set for parameter optimization. The statistics of the corpus used in this study is shown in Table 2.

Table 2. Detailed data set statistics.

Item	Training set, n	Development set, n	Test set, n
Document	149	14	117
Sentence	770	71	644
FM[a]: overall	1128	94	—[b]
FM: NA[c]	631	55	—
FM: maternal	272	24	—
FM: paternal	225	15	—
OB[d]	1439	127	—
LS[e]	596	52	—
FM-OB: overall	1064	97	—
FM-OB: NA-OB	575	57	—
FM-OB: maternal-OB	265	23	—
FM-OB: paternal-OB	224	17	—
FM-LS: overall	605	53	—
FM-LS: NA-LS	334	29	—
FM-LS: maternal-LS	145	12	—
FM-LS: paternal-LS	126	12	—

[a]FM: family member.

[b]Not available.

[c]NA: not applicable.

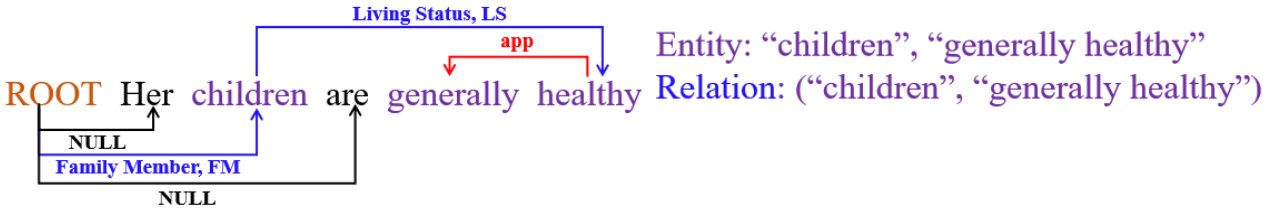
[d]OB: observation.

[e]LS: living status.

Graph-Based Schema

Similar to the dependency parsing task where each token has a head token, we transformed the family history information extraction task to a dependency parsing problem, where a dummy root (denoted by "ROOT") was appended to each sentence at the beginning and arcs denoted links between two tokens. In the "dependency parsing tree" of a sentence, tokens in each entity were connected together by an "app" arc from right to left, two entities with a relation were connected through linking the right most token by an arc labeled with the entity type, and tokens not in any entity were connected with the "ROOT" node by "NULL" arcs. Figure 1 shows an example of using a "dependency parsing tree" to represent family history information extraction, where the family member entity "children" was determined by the "Family Member" arc from "ROOT" to "children," the living status entity "generally healthy" was determined by "generally generally," and the relation between "children" and "generally healthy" was determined by the arc from "children" to "healthy" .

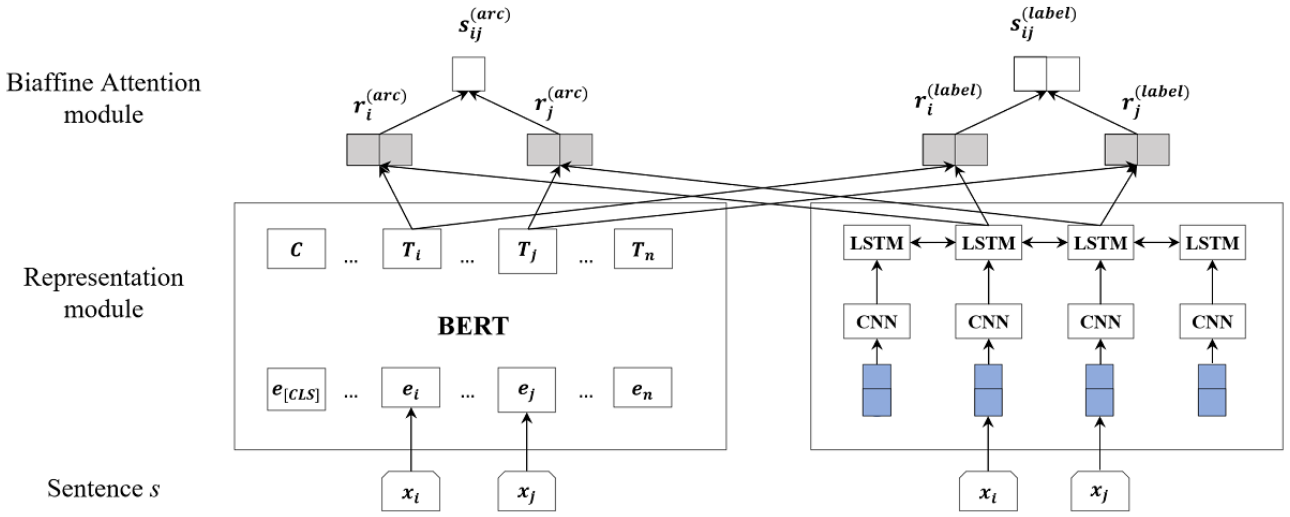
Figure 1. Example of using a graph-based schema to represent family history information.



Model Architecture

As shown in Figure 2, our model contained the following two main parts: (1) a representation module, which represented input text using BERT and CNN-BiLSTM and (2) a biaffine attention module to predict label score vectors, including unlabeled arc prediction (top left in Figure 2) and arc label prediction (top right in Figure 2). We have presented them in the following sections in detail.

Figure 2. Overview architecture of our model.



Representation Layer

Given a sentence $s = x_1 \ldots x_i \ldots x_n$, where x_i is the ith token of s, we used BERT and CNN-BiLSTM to represent it separately as follows:

where CNN [4] is first used to get the character-level representation of each token, and BiLSTM is then used to get the contextual representation of each token in CNN-BiLSTM. The final representation of token x_i is

Biaffine Attention Layer

Unlabeled Arc Prediction

Considering the ith token and the jth token, we fed their corresponding representations into a bilinear transformation extension called a biaffine function to get the score of the arc from token i (head) to j (dependent) as follows:

where $r_j^{(arc-dep)} \in R^p$ and $r_j^{(arc-head)} \in R^p$ are the outputs of multilayer perceptron, $U^{(arc)} \in R^{p \times p}$ is a weight matrix controlling the strength of the arc from token i to j, and $u^{(arc)} \in R^p$ is a bias vector.

Assume that $s_j^{(arc)} = [s_{1j}^{(arc)}; \ldots; s_{nj}^{(arc)}]$ is the score vector of all possible heads of the jth token. We adopted the softmax function to compute the probability distribution d_j of all possible heads of token j and the cross-entropy between the predicted d_j and gold standard $d_j^{(arc)}$ as the loss function as follows:

Thereafter, the best head of token j was determined according to

Arc Label Prediction

For each unlabeled arc, we need to determine its label. Assume that $s_{ij}^{(lab)} \in R^{|L|}$ is the label score vector for each arc from token i to j, where $|L|$ is the size of the label set. We can compute $s_{ij}^{(lab)}$ as follows:

where $r_j^{(label-dep)} \in R^{|L| \times p}$ and $r_j^{(label-head)} \in R^{|L| \times p}$ are outputs of the multilayer perceptron, $U^{(label)} \in R^{|L| \times p \times p}$ is a third-order tensor, $W^{(label)} \in R^{|L| \times 2p}$ is a weight matrix, and $u^{(label)} \in R^{|L|}$ is a bias vector.

We also adopted the softmax function to compute the probability distribution d_{ij} of all possible labels of the arc from token i to j and the cross-entropy between the predicted d_{ij} and gold standard $d_{ij}^{(label)}$ as the loss function as follows:

Thereafter, the best label of the arc from token i to j was determined by

The total loss function was set as

Postprocessing Rules

We designed a rule-based postprocessing module to adjust the outputs of our model. It included the following five parts:

1. Converting the output to entities and relations.

(1) Combining all tokens connected by “app” arcs to form entities and assigning them the label of their last token.

(2) If there was an arc between two entities, but not an "app" arc, there was a relation between them.

2. Normalizing family members.

(1) Converting family member entities into normalized forms as shown in Table 1. For example, we converted the recognized "father's father" into "grandfather" and "aunt's son" into "cousin."

(2) Excluding unnecessary family members. For example, a patient's nonblood relatives, such as "father" in section "partner's father," should be removed. If the family member "father" belonged to section "partner's father," we removed "father" since father-in-law was not in Table 1.

3. Determining the side of family members when using the strategy of three types of entities.

(1) If a family member was a first-degree relative, the side of the family was set as "NA."

(2) If a family member was in the section "maternal family history" or "paternal family history," the side of the family was set as maternal or paternal.

(3) If there was an indicator ("maternal" or "paternal") near a family member, the side of the family was determined by the indicator.

(4) Otherwise, the side of the family of a family member was set as "NA."

4. Determining the living status score of family members following the work of Shi et al [17].

(1) Determining the scores of the properties "Alive" and "Healthy" of a family member through searching the keywords listed in Table 3 from the family member's living status. If a living status entity contained some keywords listed in Table 2, we assigned its property scores with the corresponding scores; otherwise, both its alive score and healthy score were set as NA=1.

(2) The total living status score was determined according to the alive score and healthy score. For a relative with "Alive=Yes" and "Healthy=Yes," for example, the living status score should be 4.

5. Determining the negation information of observations.

(1) Determining the negation information of an observation through searching keywords (no, never, not, none, negative, neither, nor, unremarkable, and deny) from the observation's context. If the context of an observation contained a keyword mentioned above, we set its negation information as "Negated;" otherwise, it was set as "Non_Negated."

(2) Reversing the negation information of an observation if there were specific phrases, such as "apart from" and "except for," in the observation's context. For example, the negation information of the observation entity "Meniere disease" in "there is no history of hearing loss apart from the father's history of Meniere disease" was set as "Non_Negated" rather than "Negated."

Table 3. Keywords used to determine the properties "Alive" and "Healthy."

Property	Keywords
Alive: Yes=2	Alive and living
Alive: No=0	Dead, die, deceased, death, died, stillborn, and passed away
Healthy: Yes=2	Good, health, without problems, healthy, and well

Experimental Settings

The hyperparameters used in our experiments are listed in Table 4, and all other parameters were optimized in the validation set. The pretrained BERT model we used was [BERT-Base, Uncased] [24].

We first investigated our model in the following two settings: (1) a pipeline model that tackled unlabeled arc prediction and arc label prediction separately and (2) a joint model that tackled unlabeled arc prediction and arc label prediction simultaneously. The joint model predicated the arc and label of each token in our model jointly. The pipeline model first trained one model to predict the head of each token and then trained another model to predict the head of each token according to the result of the predicted head. Thereafter, we compared our model with the BERT-based model using the same architecture as that of the model by Shi et al [17], except that we used BERT instead of word embeddings in the input layer (denoted by BERT-2BiLSTM). Finally, we looked into the effect of the sentence representation based on CNN-BiLSTM on our model and the effect of different data sets on our model. The performance of all models for the two subtasks was measured by precision, recall, and F1 score (F1) as follows:

where TP denotes the number of true-positive samples, FP denotes the number of false-positive samples, and FN denotes the number of false-negative samples. We used the tool provided by the organizers [25] to calculate them. The tool accepted partial matching of the observations, for example, the recognized observation "diabetes" whose gold standard observation is "type 2 diabetes" was considered as a true-positive sample. The source code is available at GitHub [26].

Table 4. Major hyperparameters.

Parameter	Value
BiLSTM[a] size	256
Arc MLP[b] size	500
Label MLP size	100
BERT[c] size	768
Char embedding size	25
CNN[d] kernel size	(3, 4, 5)
Char-level CNN size	50
Dropout	0.5
Optimizer	Adam
Learning rate	2e-5
Batch size	32
Max epoch	100

[a]BiLSTM: Bidirectional Long Short Term Memory network.

[b]MLP: multilayer perceptron.

[c]BERT: Bidirectional Encoder Representation from Transformers.

[d]CNN: convolutional neural network.

Results

As shown in Table 5, the performance of the model considering five types of entities was better than that considering three types of entities. The joint model considering five types of entities achieved the highest F1 score of 0.8823 on the NER subtask and 0.7048 on the RE subtask, which were higher than the values for the joint model considering three types of entities by 1.20% on the NER subtask and 1.87% on the RE subtask.

Compared to the pipeline model, the joint model performed better on both the NER and RE. For example, when considering five types of entities, the joint model outperformed the pipeline model by 1.21% in the F1 score on the NER subtask and 1.97% in the F1 score on the RE subtask. It indicated that error propagation was partially alleviated in our joint model. When considering five types of entities, the joint model achieved higher F1 scores than BERT-2BiLSTM on the NER subtask and RE subtask by 1.18% and 0.39%, respectively.

Table 5. Performance of different models.

Subtask	Model	Three types of entities			Five types of entities		
		Precision	Recall	F1 score	Precision	Recall	F1 score
NER[a]	Pipeline	0.9254	0.8062	0.8617	0.9241	0.8223	0.8702
NER	Joint	0.9012	0.8415	0.8703	0.9154	0.8514	0.8823
NER	BERT[b]-2BiLSTM[c]	—[d]	—	—	0.9096	0.8347	0.8705
RE[e]	Pipeline	0.7909	0.6005	0.6827	0.7895	0.6051	0.6851
RE	Joint	0.7679	0.6200	0.6861	0.7717	0.6487	0.7048
RE	BERT-2BiLSTM	—	—	—	0.7686	0.6441	0.7009

[a]NER: named entity recognition.

[b]BERT: Bidirectional Encoder Representation from Transformers.

[c]BiLSTM: Bidirectional Long Short Term Memory network.

[d]Not available.

[e]RE: relation extraction.

The performance of our best model on each type of family member information and relation (except living status not provided in the test set) is listed in Table 6. On the NER subtask, our model performed better on observations than family members by 3.80% in terms of the F1 score. Among the three types of family members, our model achieved the highest F1 score of 0.8702 for maternal family member and the lowest F1 score of 0.8411 for paternal family member. On the RE subtask, the F1 score of our model on the family member-living status relation was nearly the same as that of our model on the family

member-observation relation. Among the family member-observation relations, our model performed worse on the maternal-observation relation than the other two types of relations. Among the family member-living status relations, our model performed worse on the paternal-living status relation than the other two types of relations.

Table 6. Performance of the best model on each type of family member information.

Subtask	Type	Precision	Recall	F1 score
NER[a]	FM[b]: overall	0.8814	0.8386	0.8594
NER	FM: NA[c]	0.8699	0.8515	0.8606
NER	FM: maternal	0.9185	0.8267	0.8702
NER	FM: paternal	0.8738	0.8108	0.8411
NER	OB[d]	0.9385	0.8598	0.8974
NER	LS[e]	—[f]	—	—
NER	Overall	0.9154	0.8514	0.8823
RE[g]	FM-OB: overall	0.7843	0.6397	0.7047
RE	FM-OB: NA-OB	0.8595	0.6098	0.7134
RE	FM-OB: maternal-OB	0.7067	0.6601	0.6826
RE	FM-OB: paternal-OB	0.7077	0.7150	0.7113
RE	FM-LS: overall	0.7627	0.6553	0.7050
RE	FM-LS: NA-LS	0.7627	0.6553	0.7050
RE	FM-LS: maternal-LS	0.7108	0.7375	0.7239
RE	FM-LS: paternal-LS	0.6825	0.6825	0.6825
RE	Overall	0.7717	0.6487	0.7048

[a]NER: named entity recognition.

[b]FM: family member.

[c]NA: not applicable.

[d]OB: observation.

[e]LS: living status.

[f]Not available.

[g]RE: relation extraction.

As shown in Table 7, without using the additional data for BioCreative/OHNLP 2018, our model considering five types of entities achieved an F1 score of 0.8648 on the NER subtask and 0.6612 on the RE subtask (the F1 score was significantly reduced both on the NER subtask and RE subtask), showing the importance of the data.

Table 7. Performance of our model with different data.

Subtask	Data set	Three types of entities			Five types of entities		
		Precision	Recall	F1 score	Precision	Recall	F1 score
NER[a]	2019	0.8767	0.8409	0.8584	0.8847	0.8458	0.8648
NER	2018+2019[b]	—[c]	—	—	0.9154	0.8372	0.8745
NER	2018+2019[d]	0.9012	0.8415	0.8703	0.9154	0.8514	0.8823
RE[e]	2019	0.7240	0.5973	0.6545	0.7270	0.6064	0.6612
RE	2018+2019[b]	—	—	—	0.7459	0.6265	0.6810
RE	2018+2019[d]	0.7679	0.6200	0.6861	0.7717	0.6487	0.7048

[a]NER: named entity recognition.

[b]2018+2019: the challenge submission performances of our model.

[c]Not available.

[d]2018+2019: the performances of our best model after challenge.

[e]RE: relation extraction.

Discussion

Effect of Sentence Representation

In order to investigate the effect of sentence representation based on CNN-BiLSTM on our model, we evaluated the model without using the representation and obtained an F1 score of 0.8802 on the NER subtask and an F1 score of 0.7059 on the RE subtask when considering five types of entities. The sentence representation based on CNN-BiLSTM can bring improvement in the NER subtask, but a little loss in the RE subtask. Possibly, we can only share BERT on NER and RE for further improvement.

Impact of Different Decoders on the NER Subtask

Traditional approaches regarded the NER task as a sequence labeling task, in which each token was assigned with a combined label of entity boundary and type. The entity boundaries were represented by the BIO schema, where "B" indicates the beginning of an entity, "I" indicates the inside of an entity, and "O" indicates the outside of an entity. Using a graph schema, we can also convert NER into a graph in the following way: (1) connect all tokens with "ROOT," that is, the heads of all tokens are set to 0 and (2) set the label of the nonentity token to "NULL," set the label of the last token in the entity to the entity type, and set the label of the remaining token in the entity to "app."

We compared different decoders, that is, CRF for sequence labeling, biaffine for NER only (biaffine-NER), and biaffine for joint NER and RE (biaffine-Joint). As shown in Table 8, the performance of biaffine-NER was slightly better than that of CRF, while biaffine-Joint was considerably better than the other two models. Although the head prediction was not directly related to the NER task, the arcs of different types among tokens provided global information that was beneficial to the NER task.

Table 8. Comparison of different decoders on the named entity recognition subtask.

Decoder	Three types of entities			Five types of entities		
	Precision	Recall	F1 score	Precision	Recall	F1 score
CRF[a]	0.8989	0.8316	0.8639	0.9070	0.8390	0.8717
Biaffine-NER[b]	0.9001	0.8310	0.8641	0.8895	0.8570	0.8729
Biaffine-Joint	0.9012	0.8415	0.8703	0.9154	0.8514	0.8823

[a]CRF: conditional random field.

[b]NER: named entity recognition.

Error Analysis

We performed error analysis on our model considering five types of entities in the development data set. In the case of the NER subtask, 88.24% of errors were boundary errors because of wrong "app" arc prediction, while the remaining 11.76% of errors were type errors that have a correct boundary but wrong entity type. For example, in the sentence "The paternal grandmother, age 53, has wind sucking attributed to not having intestinal during her life," the paternal entity "grandmother" with the observation entity "wind sucking" was wrongly recognized as a family member entity. In the RE subtask, all errors were caused by incorrect entities. For example, in the sentence "The patient's father is 43 years old and healthy. His father is 72 years old and was diagnosed with esophageal cancer at age 70," the family member entity "grandfather" with the observation entity "esophageal cancer" was wrongly extracted as the family member entity "father" with the observation entity

"esophageal cancer" as our model could not understand that "his" refers to "the patient's father," which needs strong indirect relative reasoning.

Limitations and Future Work

The rule-based postprocessing module in our system cannot handle all cases properly, as shown by the example in the error analysis section. In future work, we will try to solve indirect relative reasoning for further improvement.

Conclusions

In this study, we proposed a novel graph-based model with biaffine attention, where a graph-based schema was design to represent entities and relations regarding family history in a unified way and deep biaffine attention was adopted to extract the entities and relations from clinical text. Our system based on the proposed model achieved the highest F1 score of the challenge to date.

Acknowledgments

This paper was supported in part by the following 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. Lafferty J, Mccallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. 2001 Presented at: Eighteenth International Conference on Machine Learning; June 28-July 1, 2001; Williamstown, MA, USA p. 282-289.
2. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Med Inform Decis Mak 2013;13(Suppl 1):S1. [doi: 10.1186/1472-6947-13-s1-s1]
3. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv. 2015. URL: http://arxiv.org/abs/1508.01991 [accessed 2021-03-29]
4. Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; 2016; Berlin, Germany p. 1064-1074. [doi: 10.18653/v1/p16-1101]
5. Tang B, Hu J, Wang X, Chen Q. Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF. Wireless Communications and Mobile Computing 2018;2018:1-8. [doi: 10.1155/2018/2379208]
6. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273-297. [doi: 10.1007/bf00994018]
7. Sahu S, Anand A, Oruganty K. Relation extraction from clinical texts using domain invariant convolutional neural network. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. 2016 Presented at: 15th Workshop on Biomedical Natural Language Processing; 2016; Berlin, Germany p. 206-215. [doi: 10.18653/v1/w16-2928]
8. Luo Y. Recurrent neural networks for classifying relations in clinical notes. J Biomed Inform 2017 Aug;72:85-95 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.006] [Medline: 28694119]
9. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]
10. 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 Sep 01;20(5):828-835 [FREE Full text] [doi: 10.1136/amiajnl-2013-001635] [Medline: 23571849]
11. 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]
12. Zhang Y, Wang J, Tang B. UTH_CCB: A report for SemEval 2014 – Task 7 Analysis of Clinical Text. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). 2014 Presented at: 8th International Workshop on Semantic Evaluation (SemEval 2014); 2014; Dublin, Ireland p. 802-806. [doi: 10.3115/v1/S14-2142]
13. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. SemEval-2015 Task 6: Clinical TempEval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015 Presented at: 9th International Workshop on Semantic Evaluation (SemEval 2015); June 4-5, 2015; Denver, Colorado p. 806-814. [doi: 10.18653/v1/s15-2136]

14. Kelly L, Goeuriot L, Suominen H, Névéol A, Palotti J, Zuccon G. Overview of the CLEF eHealth Evaluation Lab 2016. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2016. Lecture Notes in Computer Science, vol 9822. Cham: Springer; 2016:255-266.
15. Suominen H, Salantera S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: Information Access Evaluation. Multilinguality, Multimodality, and Visualization. CLEF 2013. Lecture Notes in Computer Science, vol 8138. Berlin, Heidelberg: Springer; 2013:212-231.
16. Goeuriot L, Kelly L, Li W, Palotti J, Pecina P, Zuccon G, et al. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In: CLEF 2014 Online Working Notes. 2014 Presented at: CEUR Workshop; September 15-18, 2014; Sheffield, UK p. 43-61.
17. 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]
18. Li Q, Ji H. Incremental Joint Extraction of Entity Mentions and Relations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014 Presented at: 52nd Annual Meeting of the Association for Computational Linguistics; June 22-27, 2014; Baltimore, MD, USA p. 402-412. [doi: 10.3115/v1/p14-1038]
19. Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; August 7-12, 2016; Berlin, Germany p. 1105-1116. [doi: 10.18653/v1/p16-1105]
20. Liu S, Mojarad MR, Wang Y, Wang L. Overview of the BioCreative/OHNLP 2018 Family History Extraction Task. In: Proceedings of the BioCreative Workshop. 2018 Presented at: BioCreative Workshop; 2018; Washington, DC, USA.
21. 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: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); June 2-7, 2019; Minneapolis, MN, USA p. 4171-4186. [doi: 10.18653/v1/n19-1423]
22. Dozat T, Manning C. Deep Biaffine Attention for Neural Dependency Parsing. In: Proceedings of the 5th International Conference on Learning Representations. 2017 Presented at: 5th International Conference on Learning Representations; April 24-26, 2017; Toulon, France.
23. Yan H, Qiu X, Huang X. A Graph-based Model for Joint Chinese Word Segmentation and Dependency Parsing. Transactions of the Association for Computational Linguistics 2020 Dec;8:78-92. [doi: 10.1162/tacl_a_00301]
24. TensorFlow code and pre-trained models for BERT. GitHub. URL: https://github.com/google-research/bert [accessed 2021-03-05]
25. OHNLP/BioCreative 2018 Task 1: Family History Extraction. GitHub. URL: https://github.com/ohnlp/fh_eval [accessed 2021-03-05]
26. 2019 n2c2/OHNLP Track2 family history extraction. GitHub. URL: https://github.com/zkczjj/2019_n2c2_FHExtraction [accessed 2021-03-26]

Abbreviations

BERT: Bidirectional Encoder Representation from Transformers
BiLSTM: Bidirectional Long Short Term Memory
CNN: convolutional neural network
CRF: conditional random field
EHR: electronic health record
NER: named entity recognition
NLP: natural language processing
RE: relation extraction

Edited by Y Wang, F Shen; submitted 17.08.20; peer-reviewed by I Mircheva, X Yang; comments to author 12.10.20; revised version received 18.01.21; accepted 09.02.21; published 21.04.21.

Please cite as:
Zhan K, Peng W, Xiong Y, Fu H, Chen Q, Wang X, Tang B
Novel Graph-Based Model With Biaffine Attention for Family History Extraction From Clinical Text: Modeling Study
JMIR Med Inform 2021;9(4):e23587
URL: https://medinform.jmir.org/2021/4/e23587
doi:10.2196/23587
PMID:33881405

Original Paper

A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System

Youngjun Kim[1], PhD; Paul M Heider[1], PhD; Isabel RH Lally[2]; Stéphane M Meystre[1], PhD, MD

[1]Biomedical Informatics Center, Medical University of South Carolina, Charleston, SC, United States

[2]Department of Computer Science, College of Charleston, Charleston, SC, United States

Corresponding Author:
Youngjun Kim, PhD
Biomedical Informatics Center
Medical University of South Carolina
22 WestEdge Street Suite 200/Room WG213
Charleston, SC, 29403
United States
Phone: 1 843 792 4268
Email: youngjun.kim.res@gmail.com

Abstract

Background: Family history information is important to assess the risk of inherited medical conditions. Natural language processing has the potential to extract this information from unstructured free-text notes to improve patient care and decision making. We describe the end-to-end information extraction system the Medical University of South Carolina team developed when participating in the 2019 National Natural Language Processing Clinical Challenge (n2c2)/Open Health Natural Language Processing (OHNLP) shared task.

Objective: This task involves identifying mentions of family members and observations in electronic health record text notes and recognizing the 2 types of relations (family member-living status relations and family member-observation relations). Our system aims to achieve a high level of performance by integrating heuristics and advanced information extraction methods. Our efforts also include improving the performance of 2 subtasks by exploiting additional labeled data and clinical text-based embedding models.

Methods: We present a hybrid method that combines machine learning and rule-based approaches. We implemented an end-to-end system with multiple information extraction and attribute classification components. For entity identification, we trained bidirectional long short-term memory deep learning models. These models incorporated static word embeddings and context-dependent embeddings. We created a voting ensemble that combined the predictions of all individual models. For relation extraction, we trained 2 relation extraction models. The first model determined the living status of each family member. The second model identified observations associated with each family member. We implemented online gradient descent models to extract related entity pairs. As part of postchallenge efforts, we used the BioCreative/OHNLP 2018 corpus and trained new models with the union of these 2 datasets. We also pretrained language models using clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database.

Results: The voting ensemble achieved better performance than individual classifiers. In the entity identification task, our top-performing system reached a precision of 78.90% and a recall of 83.84%. Our natural language processing system for entity identification took 3rd place out of 17 teams in the challenge. We ranked 4th out of 9 teams in the relation extraction task. Our system substantially benefited from the combination of the 2 datasets. Compared to our official submission with F_1 scores of 81.30% and 64.94% for entity identification and relation extraction, respectively, the revised system yielded significantly better performance (P<.05) with F_1 scores of 86.02% and 72.48%, respectively.

Conclusions: We demonstrated that a hybrid model could be used to successfully extract family history information recorded in unstructured free-text notes. In this study, our approach to entity identification as a sequence labeling problem produced satisfactory results. Our postchallenge efforts significantly improved performance by leveraging additional labeled data and using word vector representations learned from large collections of clinical notes.

(JMIR Med Inform 2021;9(4):e22797) doi:10.2196/22797

KEYWORDS

natural language processing; machine learning; deep learning; named entity recognition; clinical entity identification; relation extraction

Introduction

Family history (FH) information included in the electronic health record (EHR) is important to assess the risk of inherited medical conditions. For certain diseases such as breast cancer [1,2] and colorectal cancer [3,4], FH is an important risk factor. FH information has been recorded in both structured and narrative free text, but often documented only in the latter. Polubriaginof et al [5] reported that free-text notes contained more comprehensive information than structured data. Natural language processing (NLP) has the potential to extract this information from unstructured free-text notes to improve patient care and decision making.

This manuscript describes the end-to-end information extraction (IE) system the Medical University of South Carolina (MUSC) team developed when participating in the 2019 National Natural Language Processing Clinical Challenge (n2c2)/Open Health Natural Language Processing (OHNLP) track on FH extraction [6]. This shared task is built on the BioCreative/OHNLP 2018 FH extraction task [7]. It involves (1) identifying mentions of family members and observations in EHR text notes and (2) recognizing the relations between family members, observations, and living status.

Entity identification and relation extraction are often considered subtasks of IE. The semantic types of concepts of interest have been defined for different target tasks. Named entity recognition (NER) was introduced in the sixth of a series of Message Understanding Conferences [8] and Automatic Content Extraction programs [9]. The goal of NER is to extract and classify proper named or specialized entities into predefined categories [8]. Relation extraction deals with a pair of concepts [10] (ie, binary relations) or higher-order relations, which are *n*-ary relations among *n* typed entities [11]. It aims to determine whether entities are in a relation and how they are semantically related. Medical concept extraction is closely related to our target task and has advanced from the general text NER by sharing the algorithms and features. It has aimed to extract medical information such as disease diagnoses, medications, laboratory data, and appliances from EHR text notes [12-16].

Several studies focusing on FH information have been reported. Goryachev et al [17] created a rule-based system for identifying family members and their related diagnoses. They observed that FH was often mentioned intermixed with the patient's own medical history, making this task challenging. Bill et al [18] developed an NLP system for extracting FH information from History and Physical notes. Their NLP pipeline identified family member and observation entities, relations between them, and attributes such as vital status and age. FH information extraction was the focus of the BioCreative/OHNLP 2018 task [7]. The best performance on this shared task was achieved by Shi et al [19] with F_1 scores of 89.01% on subtask 1 and 63.59% on subtask 2. They proposed joint modeling of entities and relations by 2 stacked neural networks with shared parameters.

The goal of this study was to extract the health information of patients and their relatives from unstructured EHR notes. Our system aims to achieve a high level of performance in this task by integrating heuristics and advanced information extraction methods. We approach entity identification as a sequence labeling problem. We applied a bidirectional long short-term memory (Bi-LSTM) [20] algorithm, a widely used structured prediction algorithm. The input of the LSTM network included vector representations generated by Embeddings from Language Models (ELMo) [21] contextual embeddings. We hypothesized that applying the LSTM to this problem can yield accurate FH information extraction. Our voting ensemble is created based on the fact that the LSTM algorithm is not deterministic; that is, every time the model is trained, the results vary. The proposed ensemble can provide efficient and convenient integration of individual LSTM models. For relation extraction, we implemented online gradient descent (OGD) [22] models with lexical features.

This study's contribution also includes improved performance on both subtasks by exploiting additional labeled data and clinical text–based embedding models. We added other labeled data used in the previous shared task to the training set. We retrained the classifier using a larger set of training data. We also used word embeddings pretrained with large quantities of clinical text. Our experimental results show that these efforts significantly improve the performance of both subtasks, especially relation extraction.

The following sections describe the details of the 2 subtasks and discuss IE models developed to recognize the entities and their relations from EHRs. We then present the experimental results and investigate the performance improvements resulting from our postchallenge efforts.

Methods

Our research focuses on the extraction of mentions of family members and related information recorded in EHR text notes. The first subtask, entity identification, involves detecting 2 types of entities: family members and observations. Only relatives in the first degree (eg, '*Mother*' and '*Son*') and second degree (eg, '*Grandparent*' and '*Cousin*') are annotated [7]. Other relatives such as '*Spouse*' and '*Nephew*' are excluded. The normalized name and the side of family are annotated as attributes of each family member. Observation (disease) entities in the family history are also annotated. The second subtask, relation extraction, is to determine the existence of relations between family members and other information (ie, living status or observation). Two types of relations were therefore annotated: family member-living status and family member-observation. For relations between a family member and living status, the

score representing the health status of the family member is annotated. Negation information is annotated to indicate whether the observation is negated in the relation between a family member and the associated observation.

Data Description

Clinical text notes representing patient FH information were selected from the Mayo Employee and Community Health cohort [7]. Table 1 shows the number of annotated entities and relations in the training set. The training set includes 99 clinical notes with 801 family member and 978 observation entities. Living status entities are less common and account for about half of the number of family members. For the observation category, the number of relations is less than the number of entities. This means that some observations are not related to any family member.

Table 1. Number of annotated entities and relations in the training set.

Variable	Entities	Relations
Family member	801	N/A[a]
Living status	415	425
Observation	978	753

[a]N/A: not applicable because relations between family members were not annotated.

Entity Identification Methods

We addressed entity identification with rule-based and machine learning–based approaches. We describe each approach and present a voting ensemble–based method.

Rule-Based System for Family Member Entities

Our rule-based system for family member entity recognition uses a sliding window with simple term matching and part-of-speech filtering. We used NLTK [23] (a Python Natural Language Toolkit) to split each note into sentences and then each sentence into tokens annotated with part-of-speech tags. Each token matching a relevant family member term (eg, "*daughter*", "*son*", or "*child*") that was also tagged as a noun (ie, NN, NNP, or NNS) was flagged as a valid mention.

Machine Learning–Based Models

We trained sequence labeling models using Bi-LSTM [20,24] to assign a semantic category label to each word in a sequence. Bi-LSTM can combine both forward and backward information of each word.

For this sequence labelling problem, we tokenized the input text. The training data were annotated with BIO token tags (B: beginning, I: inside, or O: outside of an entity; eg, "B-observation" for a token at the beginning of an observation mention). We also included the outputs of the 2 external resources (the 2010 Informatics for Integrating Biology and the Bedside [i2b2] [25] and MetaMapLite [26]) described in the following paragraphs as inputs to the LSTM network. Similar to the word token, the prediction from each external resource was also encoded with BIO tags.

First, we used the medical concept extraction model trained with the 2010 i2b2 challenge data [25]. The training set containing 349 text documents was used to create a Bi-LSTM model that identified medical *problem*, *treatment*, and *test* concepts from the FH extraction task corpus. We also used MetaMapLite [26] (2019 AA version) to identify Unified Medical Language System (UMLS) Metathesaurus concept mentions along with their semantic type. We aligned MetaMap outputs with the entity types of subtask 1 to choose the relevant semantic types. Table 2 lists the 10 most frequently aligned UMLS semantic types used by MetaMap for observation entity extraction. The first and second columns display semantic type names and abbreviations. The third column shows the number of observation entities from the training corpus aligned with each semantic type. The last column shows the mapping probability for each semantic type and observation category. For instance, "Disease or Syndrome" was mapped to the observation category with a probability of 79.89%. We used the training data to automatically create these heuristics. We used all (21) semantic types with a mapping probability of over 70%. The output semantic type was converted to a family member or observation entity, such as B-family_member or I-observation.

Our Bi-LSTM model incorporated 2 embedding layers for pretrained word embeddings. We used dependency-based embeddings by Komninos and Manandhar [27] as static word embeddings. These embeddings were trained using the structure of dependency graphs. They were built with the English Wikipedia Dump of August 2015. As context-dependent embeddings, we used the ELMo [21] model trained on a dataset of 5.5 billion tokens from Wikipedia and the news crawl corpus. The output of each external resource (the 2010 i2b2 and MetaMapLite) was represented as a one-hot vector and mapped to a 10-dimensional embedding. The concatenation of these embeddings (2 pretrained embeddings and 2 one-hot vectors) was fed to the LSTM layer.

To fine-tune the parameters of LSTM models, we randomly selected 10 documents from the training set (about 10% of the training set) as held-out data. We tuned the hyperparameters to maximize the F_1 score with the held-out data. After experimenting with different dropout [28] rates of 10%, 20%, 30%, 40%, and 50%, the models were trained using the Nadam [29] optimizer for 30 epochs with a dropout rate of 50%.

Table 2. The 10 most frequent Unified Medical Language System (UMLS) semantic types aligned with labeled observations in the training set.

Semantic type name	Abbreviation	Count	Probability, %
Disease or syndrome	dsyn	433	79.89
Neoplastic process	neop	165	78.20
Mental or behavioral dysfunction	mobd	59	74.68
Sign or symptom	sosy	28	70.00
Congenital abnormality	cgab	27	90.00
Anatomical abnormality	anab	10	83.33
Body system	bdsy	8	72.73
Tissue	tisu	7	100.00
Cell	cell	5	83.33
Physiologic function	phsf	4	80.00

We trained 10 different Bi-LSTM models that use the same hyperparameters but differ in random weight initialization and shuffling of training data. Then, we created a voting ensemble method that combined the predictions of all Bi-LSTM trials. Although these LSTM models were trained with the same hyperparameters, we hypothesized that they can be contributory to the voting ensemble in terms of diversity. Reimers and Gurevych [30] showed that nondeterministic LSTMs can even lead to statistically significant differences between multiple runs.

The voting ensemble collected candidate entities that received more votes than the voting threshold. When there were overlapping text spans on 2 different entities, the entity with more votes was selected. For overlapping entities with the same vote count, the one produced by the higher-ranking model was selected. To determine the ranking of 10 individual models, we measured how each model agreed with the other 9 models. Rankings were based on F_1 scores measured with other models. The higher the average F_1 score, the higher the model ranking.

Heuristic Rules for Family Member Attributes

We assigned each family member entity a normalized form using a simple dictionary-based mapping. For example, a family member with the text "*his dad*" was assigned '*Father.*' We changed the text to lower case and removed the numeric values (eg, "*three uncles*" becomes '*Uncle*'). We also looked at the preceding words to search for another family member term that modified the target entity. When such a term was found, normalization was performed taking it into account. For example, in the phrase "*mother has sister,*" the family member '*sister*' was normalized to '*Aunt.*'

Our rule-based system looked at words in sentences near the family member and considered the degree of relatives to determine the family side. For each family member who was not a first-degree relative, the side of family (ie, '*Paternal*' or '*Maternal*') was assigned. For each label, we compiled the list of cue words indicating the side of family. For example, the cues for *Paternal* included '*paternal,*' '*patient's father,*' '*father had,*' and '*paternal family history.*' First, we searched for cue words within the entity term itself. If no cue word was found, the search was expanded to sentence boundaries.

Relation Extraction Methods

Subtask 2 aimed to identify related pairs of 3 entity types: family members, observations, and living status. Two types of relation exist between the 2 entities: family member-living status relations and family member-observation relations. We trained 2 relation extraction models. The first model determined the living status of each family member. The second model identified observations associated with each family member.

For 2 binary-class models, we defined lexical features: words contained in each concept, 7 preceding and 7 following words for each concept, and the words between the 2 concepts. We also created 1 feature to measure the number of family member entities appearing between the pair. We created 2 binary-class OGD (also called stochastic gradient descent) [22] classifiers using the Vowpal Wabbit [31] online learning library. This online learning algorithm is getting more attention recently in large-scale machine learning problems. Using the default hyperparameters, each model was trained for 100 iterations.

Training examples included positive examples (participating in a relation) and negative examples (pairs of entities that are not related to each other). Pairs of reference standard entities were used to train the classifiers. Entity pairs identified by the aforementioned voting ensemble were used as test examples. We filtered out the negative examples when there was a carriage return character ('\n') between the pair.

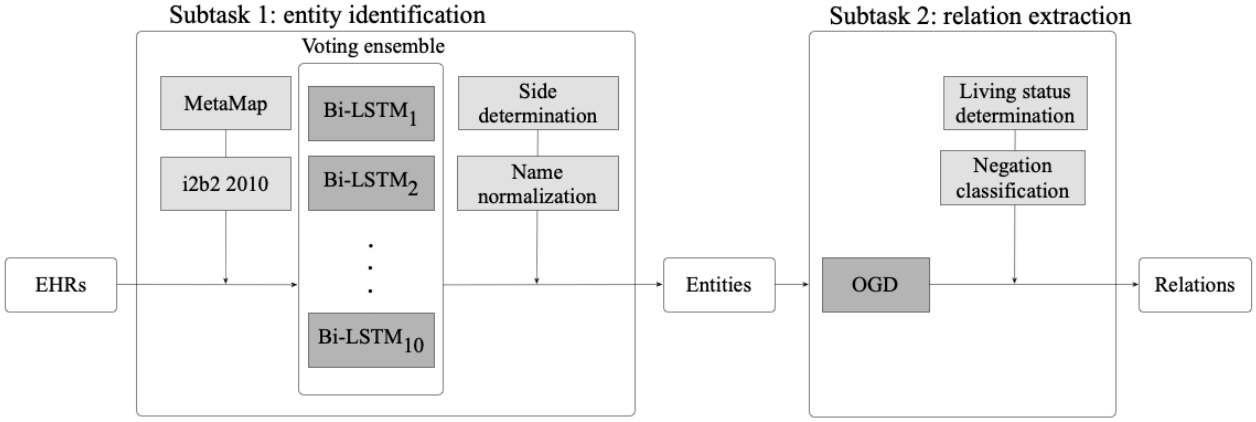
For living status relations, once we extracted phrases that represent the living status of each family member, we assigned scores for the *alive* and *healthy* attributes. We compiled *not alive* (ie, dead) and *healthy* cues from the training data and calculated the score using the text phrase of each living status entity. If our algorithm detected any trigger phrase of *not alive* (eg, "*deceased,*" "*passed away,*" and "*no longer living*"), the algorithm assigned a score of 0. Otherwise, if the family member was in good health (eg, "*good general health,*" "*healthy,*" and "*alive and well*"), the algorithm assigned a score of 4. If no cues of *not alive* or *healthy* were found, a score of 2 was assigned.

For each observation entity in the relation, we needed to determine whether it was negated or not. We used FastContext [32], an efficient and scalable Java implementation of the ConText algorithm [33] with customized trigger terms. After

manually analyzing the examples from the training data, we added new trigger terms such as "*not aware of*," "*not significant*," and "*no family history of*." For this binary classification, the algorithm detected the negated contextual attribute in the sentence for the observation entity and assigned 1 of 2 values: *Negated* or *Non_Negated*.

In summary, we built an end-to-end system with multiple IE and attribute classification components, as shown in Figure 1. The architecture includes a voting ensemble with Bi-LSTM models that accept the outputs of the MetaMap and 2010 concept models, an OGD model that extracts relations between entities, and postprocessing modules for family side, name normalization, living status, and negation classification.

Figure 1. End-to-end system architecture. Bi-LSTM: bidirectional long short-term memory; EHR: electronic health record; i2b2: Informatics for Integrating Biology and the Bedside; OGD: online gradient descent.



Improvements to Both Subtasks After the Shared Task Challenge

This subsection describes further improvements to both entity identification and relation extraction as postchallenge efforts. We made 2 major changes in the pipeline system. The first revision was the addition of labeled examples to the training data. We used another text collection created for the 2018 BioCreative/OHNLP shared task [7] to build new Bi-LSTM and OGD models. The combined dataset included the original 99 clinical notes and 50 text files used in the 2018 BioCreative/OHNLP test set. Extending from the previous models used for submission to the shared task, we investigated how well the new model trained with the union of 2 datasets performed. We trained the new models by reusing the classifier configuration optimized with the 2019 training data.

Next, we used word embeddings trained with clinical text to construct vector representations of words. We pretrained 2 language models. One was trained using fastText [34] as static word embeddings, and the other was trained using ELMo [21] contextual embeddings. We used all clinical notes from the Medical Information Mart for Intensive Care (MIMIC-III) clinical database (version 1.4) [35]. We pretrained ELMo embeddings by following the default hyperparameter setting used for other publicly available ELMo models [21]. Pretraining lasted about 3 months, and it was manually stopped after 1,073,750 iterations. This process was performed on a NVIDIA Tesla P4 GPU.

From these pretrained language models, we generated word vectors as input features. Then, we created new Bi-LSTM models for entity identification. As with the previous models, these models were trained for 30 epochs with 50% dropout to the recurrent units. Naturally, the predictions of these new Bi-LSTM models were used to create test instances that paired the 2 entities for relation extraction. In the next section, we present the experimental results from our official submission and revised systems.

Results

The input for subtask 1 (entity identification) was clinical text notes. The entity annotation file for subtask 1 contains family member and observation entities, one entity per line. The family side is provided for each family member entity. For subtask 2 (relation extraction), entity annotations were additionally used as input. The relation annotation file for task 2 contains 2 entities with their relation, 1 relation per line. Each living status relation has a score to represent living status. In each observation relation, the negation of the observation entity was identified.

Evaluation Metrics

We measured recall, precision, and F_1 score (harmonic mean of recall and precision with equal weight). We used the 2019 n2c2/OHNLP shared task [6] evaluation script to calculate performance measures. To be considered a true positive, the entity attributes must also match. For observation entities, a match was counted if the reference annotation contained 1 or more words in common with the system-detected concept.

Results for the 2019 n2c2/OHNLP Shared Task

The 2019 n2c2/OHNLP shared task corpus consisting of a test set of 117 clinical notes was used for the evaluation. First, we present the results generated by systems implemented for the 2019 n2c2/OHNLP shared task submission.

Table 3 shows the microaveraged overall precision, recall, and F_1 score for each of our submissions. The following 3 systems were submitted for subtask 1: System 1.1 was a rule-based system for collecting family member entities and a voting ensemble with a voting threshold of 5 for extracting observation

entities, system 1.2 was a voting ensemble consisting of 10 trials with a voting threshold of 5 for extracting family member and observation entities, and system 1.3 was a voting ensemble with a voting threshold of 6. Among them, system 1.2 achieved the highest F_1 score, 81.30%, in subtask 1.

Table 3. Results produced for the 2019 National Natural Language Processing Clinical Challenge (n2c2)/Open Health Natural Language Processing (OHNLP) shared task.

System	Precision score	Recall score	F_1 score
Subtask 1 (entity)			
System 1.1	72.61	86.01	78.74
System 1.2	78.90	83.84	81.30
System 1.3	80.29	81.98	81.13
Subtask 2 (relation)			
System 2.1	65.48	64.41	64.94
System 2.2	66.37	62.78	64.53
System 2.3	68.23	59.79	63.73

Similarly, we submitted 3 systems for subtask 2: System 2.1 was an OGD model with input pairs generated from predictions of the voting ensemble with a voting threshold of 4, system 2.2 was an OGD model with outputs from system 1.2, and system 2.3 was an OGD model with outputs from system 1.3. System 2.1 achieved a higher F_1 score than the others. The range of vote thresholds for task submission was selected after experimenting with values from 1 to 10 on the validation set. The highest F_1 score was obtained in subtask 2 with a voting threshold of 5 on the validation set.

Improved Results After the Shared Task

We report the results of further improvements for both subtasks as described earlier. The contributions of features or data are shown in Table 4. Systems from rows 1 to 3 were developed for the 2019 n2c2/OHNLP challenge, and rows 4 and 5 were postchallenge efforts.

Table 4. Improved performance by feature or data accumulation.

System	Precision score	Recall score	F_1 score
Subtask 1 (entity)			
(1) word	78.50	84.28	81.26
(2) + MetaMap, i2b2[a] 2010	78.87	84.34	81.48
(3) + voting	78.90	83.84	81.30
(4) + 2018 data (postchallenge)	83.63	86.69	85.13
(5) + MIMIC[b] embeddings (postchallenge, $[2018 + 2019]_{mim}$)	84.83	87.24	86.02
Subtask 2 (relation)			
(1) word	65.35	61.14	63.14
(2) + MetaMap, i2b2 2010	66.34	61.24	63.66
(3) + voting	66.37	62.78	64.53
(4) + 2018 data (postchallenge)	72.15	70.79	71.46
(5) + MIMIC embeddings (postchallenge, $[2018 + 2019]_{mim}$)	73.27	71.70	72.48

[a]i2b2: Informatics for Integrating Biology and the Bedside.

[b]MIMIC: Medical Information Mart for Intensive Care.

As a baseline, only sequences of word tokens were used as input to train the Bi-LSTM models (row 1). The system was enhanced with the output of MetaMapLite [26] and the 2010 i2b2 [25] concept model as inputs (row 2). For rows 1 and 2, we report the average value between the 10 trials of each Bi-LSTM model. From row 3, the results of applying the voting ensemble are displayed. For comparison, we report results with a voting threshold of 5. Row 4 shows a further performance improvement when the 2018 BioCreative/OHNLP shared task [7] data were added. This additional training example achieved substantial performance improvements in both subtasks. Compared to the submission for the challenge (row 3), the recall increased by 8.01% (70.79%-62.78%) in subtask 2. MIMIC embeddings (row 5) allowed for an improvement over general text embeddings. They led to F_1 scores of 86.02% and 72.48% for subtask 1 and subtask 2, respectively. We used a chi-squared

test to measure statistical significance. The significance level was set to .05. The performance of the full-featured system, called *(2018 + 2019)$_{mim}$*, (row 5) was significantly better than other systems with *P* values <.001 except the system with the 2018 BioCreative/OHNLP shared task data (row 4).

Table 5 displays the precision, recall, and F_1 scores of relation categories produced by the $(2018 + 2019)_{mim}$ system. F_1 scores for living status relations were 84.62% and 74.72% for subtask 1 and subtask 2, respectively. It was more challenging to determine whether the pair of family member and observation was related. For observation relations, the F_1 score was 71.79%, which was lower than for living status relations. A manual analysis of labeled examples from the training set revealed that distant pairs of family member and observation appeared more often than living status entities. In addition, there were more unrelated entity pairs (ie, negative examples) because many observation entities were not involved in the relation.

Table 5. Results of full-featured system for each relation category.

System	Precision score	Recall score	F_1 score
Subtask 1 (entity)			
Living status	83.08	86.21	84.62
Observation	85.99	87.92	86.94
Overall	84.83	87.24	86.02
Subtask 2 (relation)			
Living status	73.28	76.22	74.72
Observation	73.27	70.37	71.79
Overall	73.27	71.70	72.48

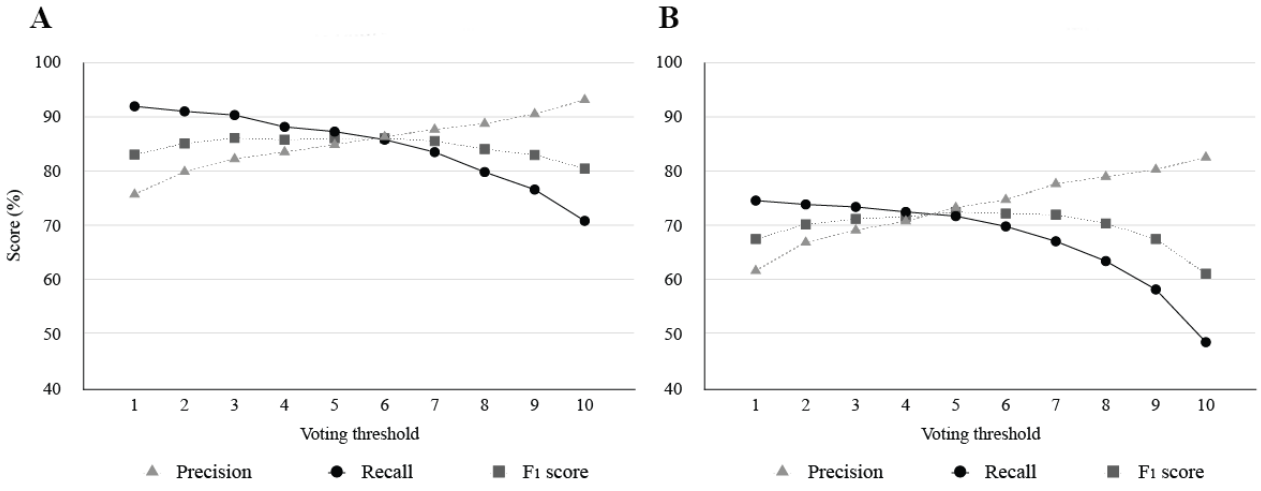
Discussion

The experimental results show that our end-to-end pipeline system substantially benefited from the combination of the 2 datasets. Another finding is that a voting ensemble could achieve better performance than individual classifiers. This section analyzes the improvements resulting from the voting ensemble method. We also describe the detailed results of attribute classification.

Voting Ensemble Analysis

We analyzed the performance of the voting ensemble at each voting threshold. Figure 2 shows the results of the voting ensembles with 5 trials of the $(2018 + 2019)_{mim}$ system. The graphs on the left and right represent the results of subtask 1 and subtask 2, respectively. The y-axis scale of each graph does not start at zero to focus on the value ranges of interest. The results with voting thresholds ranging from 1 to 10 are presented. The curves show that as the threshold gets higher, precision increases but recall simultaneously decreases. When the threshold was set to 3, the ensemble achieved the highest F_1 score (86.07%) in subtask 1. For subtask 2, the ensemble obtained an F_1 score of 72.48% at the voting threshold of 5.

Figure 2. Results of the voting ensemble for (A) subtask 1: entity identification and (B) subtask 2: relation extraction.



Attribute Classification Analysis

We applied heuristics to determine the attributes of entities. As the entity-level reference standard in the test set was being withheld, we evaluated the performance of these rule-based methods on the training set. Table 6 shows the accuracy of the 4 classification tasks with the given reference standard concepts. Accuracy was computed as the percentage of correct predictions among total instances. The accuracy of family member normalization was 94.01%. Our classifier rarely failed to assign normalized terms to some entities. For example, our dictionary did not contain normalized terms for "*twin*" and "*paternal relatives.*" Most errors occurred when the normalized term did not match the actual relationship with the patient. For example, although it said "*brother*" in the text, it sometimes referred to the relationship with the patient's parents, not the patient himself. The classifier often could not determine the family member as the patient's "*Uncle.*" This type of error was propagated in family-side decisions because the family-side information should only be provided to first-class relatives.

Table 6. Accuracy of attribute classification of given reference standard concepts.

Task	Accuracy (%)
Normalization of family members	94.01
Determination of the side of family	95.38
Assessment of family member's living status	92.53
Detection of negation information for observations	98.06

In addition to the entity-level assessment described in the previous paragraph, we conducted another document-level evaluation of the entity attributes against the test set. To measure the performance impact of each attribute classification, the system was tested by ignoring one attribute of the entity. Table 7 shows the results of the 2 subtasks on the test set by the $(2018 + 2019)_{mim}$ system. We report results for living status and negative information only for subtask 2 because they are not considered in subtask 1. A match is made if the system correctly detects an entity while the attribute is ignored. Compared to the default evaluation, which considered all attributes, it led to higher values for all metrics. Ignoring living status scores had the biggest impact. If the living status of every family member was correctly determined, the F_1 score could be increased by about 2%. Negative information had the least impact because it only applied to observations and might have been determined more accurately than other attributes.

Table 7. Performance impact of attribute classification.

System	Precision score	Recall score	F_1 score
Subtask 1 (entity)			
Default evaluation	84.83	87.24	86.02
Ignoring the side of the family	85.92	89.11	87.49
Ignoring the living status	N/A[a]	N/A[a]	N/A[a]
Ignoring negation	N/A[b]	N/A[b]	N/A[b]
Subtask 2 (relation)			
Default evaluation	73.27	71.70	72.48
Ignoring the side of the family	75.33	73.64	74.48
Ignoring the living status	75.40	73.78	74.58
Ignoring negation	73.85	72.90	73.37

[a]N/A: not applicable as the living status information was removed from evaluation for subtask 1.

[b]N/A: not applicable as the negation information was removed from evaluation for subtask 1.

Limitations

We observed in this study that determining the voting threshold can be challenging for both subtasks. Our results showed that the best performing voting ensemble for one task did not achieve the highest accuracy for the other task. More efficient ensemble approaches will be desired to provide more diversity between individual models and reduce the error rate through optimal control of agreements among them. In the relation extraction task, the negative examples were filtered out when there was a carriage return character between the pairs, because they rarely appeared in the training data (about 2.6%). This instance pruning would make it impossible to find pairs of entities that existed in different sentences but were related. When training new models by combining 2 corpora, we reused the classifier configuration optimized for the 2019 n2c2 model. New development data randomly selected from both corpora would be needed for hyperparameter tuning.

Conclusions

We presented a hybrid method that combined machine learning and rule-based approaches developed as part of the 2019 n2c2/OHNLP track on FH extraction [6]. The MUSC team ranked 3rd and 4th among the participating teams in subtask 1 and subtask 2, respectively. This study demonstrated that our end-to-end pipeline system could successfully extract FH information recorded in unstructured narrative free text. Our experimental results confirmed that the voting ensemble of multiple trials outperformed the individual classifiers that produced nondeterministic results. Our postchallenge efforts significantly improved performance by leveraging additional labeled data and using word vector representations learned from large collections of clinical notes.

Further research includes creating machine learning–based classifiers that will replace rule-based systems that determine the attributes of entities. They could lead to more accurate results on attribute classification as reported in several studies carried out for similar clinical NLP tasks [36-39]. Another direction for future work is to exploit unlabeled data to collect texts from the family history section. For efficient extension of the amount of training data, semisupervised learning can be employed with an instance selection method that uses text similarity measures to consider representativeness and diversity [40].

Acknowledgments

This research was supported in part by the National Cancer Institute (R42CA180190) and the SmartState endowment. We gratefully acknowledge the efforts and support of the 2019 n2c2 shared task organizers.

Authors' Contributions

YK and PH made substantial contributions to the design and implementation of the research and to the analysis of the experimental results. All authors drafted the work or revised it critically. YK drafted the initial manuscript. PH and SM provided critical revision of the manuscript. All authors gave final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflicts of Interest

None declared.

References

1. Colditz GA, Rosner BA, Speizer FE. Risk factors for breast cancer according to family history of breast cancer. For the Nurses' Health Study Research Group. J Natl Cancer Inst 1996 Mar 20;88(6):365-371. [doi: 10.1093/jnci/88.6.365] [Medline: 8609646]
2. 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]
3. Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Speizer FE, Willett WC. A prospective study of family history and the risk of colorectal cancer. N Engl J Med 1994 Dec 22;331(25):1669-1674. [doi: 10.1056/NEJM199412223312501] [Medline: 7969357]
4. Askling J, Dickman PW, Karlén P, Broström O, Lapidus A, Löfberg R, et al. Family history as a risk factor for colorectal cancer in inflammatory bowel disease. Gastroenterology 2001 May;120(6):1356-1362. [doi: 10.1053/gast.2001.24052] [Medline: 11313305]
5. Polubriaginof F, Tatonetti NP, Vawdrey DK. An Assessment of Family History Information Captured in an Electronic Health Record. AMIA Annu Symp Proc 2015;2015:2035-2042 [FREE Full text] [Medline: 26958303]
6. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, et al. Family History Extraction From Synthetic Clinical Narratives Using Natural Language Processing: Overview and Evaluation of a Challenge Data Set and Solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) Competition. JMIR Med Inform 2021 Jan 27;9(1):e24008 [FREE Full text] [doi: 10.2196/24008] [Medline: 33502329]
7. Liu S, Rastegar-Mojarad M, Wang Y, Wang L. Overview of the BioCreative/OHNLP 2018 Family History Extraction Task. 2018 Presented at: ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; August 29-September 1, 2018; Washington DC. [doi: 10.1145/3233547.3233672]
8. Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History. In: Proceedings of the 16th Conference on Computational Linguistics - Volume 1. 1996 Presented at: 16th Conference on Computational Linguistics; August 1996; Stroudsburg, PA p. 466-471. [doi: 10.3115/992628.992709]
9. Doddington G, Mitchell A, Przybocki M, Ramshaw L, Strassel S, Weischedel R. The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation. 2004 Presented at: Fourth International Conference on Language Resources and Evaluation (LREC-); May 2004; Lisbon, Portugal.
10. Chinchor NA. Overview of MUC-7/MET-2. 1998 Presented at: Seventh Message Understanding Conference (MUC-7); April 29-May 1, 1998; Fairfax, VA.

11. McDonald R, Pereira F, Kulick S, Winters S, Jin Y, White P. Simple algorithms for complex relation extraction with applications to biomedical IE. 2005 Presented at: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; June 2005; Ann Arbor, MI p. 491-498. [doi: 10.3115/1219840.1219901]
12. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-174 [FREE Full text] [doi: 10.1136/jamia.1994.95236146] [Medline: 7719797]
13. Christensen L, Haug P, Fiszman M. MPLUS: A Probabilistic Medical Language Understanding System. 2002 Presented at: ACL-02 Workshop on Natural Language Processing in the Biomedical Domain; July 2002; Philadelphia, PA p. 29-36. [doi: 10.3115/1118149.1118154]
14. Heinze DT, Morsch M, Sheffer R, Jimmink M, Jennings M, Morris W, et al. LifeCode: A deployed application for automated medical coding. AI Magazine 2001;22(2):76. [doi: 10.1609/aimag.v22i2.1562]
15. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6:30 [FREE Full text] [doi: 10.1186/1472-6947-6-30] [Medline: 16872495]
16. 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]
17. 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]
18. 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]
19. 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]
20. 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]
21. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. Association for Computational Linguistics; New Orleans, Louisiana. 2018. URL: https://arxiv.org/abs/1802.05365 [accessed 2021-04-07]
22. Bottou L. Online learning and stochastic approximations. In: Saad D, editor. On-line learning in neural networks. Cambridge, MA: Cambridge University Press; 1999:9-42.
23. Bird S, Loper E. NLTK: the Natural Language Toolkit. 2004 Presented at: ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics; July 2004; Barcelona, Spain p. 214-217. [doi: 10.3115/1118108.1118117]
24. Schuster M, Paliwal K. Bidirectional recurrent neural networks. IEEE Trans. Signal Process 1997;45(11):2673-2681. [doi: 10.1109/78.650093]
25. Uzuner O, 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]
26. 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]
27. Komninos A, Manandhar S. Dependency based embeddings for sentence classification tasks. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2016; San Diego, CA p. 1490-1500. [doi: 10.18653/v1/n16-1175]
28. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 2014;15(1):58. [doi: 10.5555/2627435.2670313]
29. Dozat T. Incorporating nesterov momentum into adam. 2016 Presented at: International Conference on Learning Representations (ICLR) 2016; May 2-4, 2016; San Juan, Puerto Rico.
30. Reimers N, Gurevych I. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. 2017 Presented at: Conference on Empirical Methods in Natural Language Processing; September 7–11, 2017; Copenhagen, Denmark. [doi: 10.18653/v1/d17-1035]
31. Langford J, Li L, Strehl A. Vowpal wabbit online learning project: Technical report. Machine learning and learning theory research. 2007 Dec 21. URL: https://hunch.net/?p=309 [accessed 2019-12-10]
32. Shi J, Hurdle JF. Trie-based rule processing for clinical NLP: A use-case study of n-trie, making the ConText algorithm more efficient and scalable. J Biomed Inform 2018 Sep;85:106-113 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.002] [Medline: 30092358]
33. Chapman W, Chu D, Dowling J. ConText: An algorithm for identifying contextual features from clinical text. 2007 Presented at: Workshop on BioNLP: Biological, Translational, and Clinical Language Processing; June 2007; Prague, Czech Republic p. 81-88. [doi: 10.3115/1572392.1572408]
34. 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]

35. Johnson AEW, Pollard TJ, Shen L, Lehman LH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016 May 24;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
36. Aberdeen J, Bayer S, Yeniterzi R, Wellner B, Clark C, Hanauer D, et al. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int J Med Inform 2010 Dec;79(12):849-859. [doi: 10.1016/j.ijmedinf.2010.09.007] [Medline: 20951082]
37. de Bruijn B, Cherry C, Kiritchenko S, Martin J, Zhu X. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. J Am Med Inform Assoc 2011;18(5):557-562 [FREE Full text] [doi: 10.1136/amiajnl-2011-000150] [Medline: 21565856]
38. Kim Y, Riloff E, Meystre SM. Improving classification of medical assertions in clinical notes. 2011 Presented at: 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies; June 2011; Portland, OR p. 311-316.
39. Bejan CA, Vanderwende L, Xia F, Yetisgen-Yildiz M. Assertion modeling and its role in clinical phenotype identification. J Biomed Inform 2013 Feb;46(1):68-74 [FREE Full text] [doi: 10.1016/j.jbi.2012.09.001] [Medline: 23000479]
40. Kim Y, Riloff E, Meystre SM. Exploiting Unlabeled Texts with Clustering-based Instance Selection for Medical Relation Classification. In: AMIA Annu Symp Proc. 2017 Presented at: AMIA Annual Symposium Proceedings; Nov 4-Nov 8; Washington, DC p. 1060-1069 URL: http://europepmc.org/abstract/MED/29854174

Abbreviations

Bi-LSTM: bidirectional long short-term memory
EHR: electronic health record
ELMo: embeddings from language models
FH: family history
i2b2: Informatics for Integrating Biology and the Bedside
IE: information extraction
LSTM: long short-term memory
MIMIC: Medical Information Mart for Intensive Care
MUSC: Medical University of South Carolina
n2c2: National Natural Language Processing Clinical Challenges
NER: named entity recognition
NLP: natural language processing
OGD: online gradient descent
OHNLP: Open Health Natural Language Processing
UMLS: unified medical language system

Edited by Y Wang, F Shen; submitted 30.08.20; peer-reviewed by J Zheng, R Abeysinghe, M Torii; comments to author 21.10.20; revised version received 15.12.20; accepted 19.02.21; published 22.04.21.

Please cite as:
Kim Y, Heider PM, Lally IRH, Meystre SM
A Hybrid Model for Family History Information Identification and Relation Extraction: Development and Evaluation of an End-to-End Information Extraction System
JMIR Med Inform 2021;9(4):e22797
URL: https://medinform.jmir.org/2021/4/e22797
doi:10.2196/22797
PMID:33885370

Original Paper

Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis

Maciej Rybinski[1], PhD; Xiang Dai[1,2], MSc; Sonit Singh[1,3], MSc; Sarvnaz Karimi[1], PhD; Anthony Nguyen[4], PhD

[1]Commonwealth Scientific and Industrial Research Organisation, Sydney, Australia

[2]University of Sydney, Sydney, Australia

[3]Macquarie University, Sydney, Australia

[4]Commonwealth Scientific and Industrial Research Organisation, Brisbane, Australia

Corresponding Author:
Maciej Rybinski, PhD
Commonwealth Scientific and Industrial Research Organisation
Marsfield
Sydney
Australia
Phone: 61 293724222
Email: maciek.rybinski@csiro.au

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/5/e30153

Abstract

Background: The prognosis, diagnosis, and treatment of many genetic disorders and familial diseases significantly improve if the family history (FH) of a patient is known. Such information is often written in the free text of clinical notes.

Objective: The aim of this study is to develop automated methods that enable access to FH data through natural language processing.

Methods: We performed information extraction by using transformers to extract disease mentions from notes. We also experimented with rule-based methods for extracting family member (FM) information from text and coreference resolution techniques. We evaluated different transfer learning strategies to improve the annotation of diseases. We provided a thorough error analysis of the contributing factors that affect such information extraction systems.

Results: Our experiments showed that the combination of domain-adaptive pretraining and intermediate-task pretraining achieved an F1 score of 81.63% for the extraction of diseases and FMs from notes when it was tested on a public shared task data set from the National Natural Language Processing Clinical Challenges (N2C2), providing a statistically significant improvement over the baseline (P<.001). In comparison, in the 2019 N2C2/Open Health Natural Language Processing Shared Task, the median F1 score of all 17 participating teams was 76.59%.

Conclusions: Our approach, which leverages a state-of-the-art named entity recognition model for disease mention detection coupled with a hybrid method for FM mention detection, achieved an effectiveness that was close to that of the top 3 systems participating in the 2019 N2C2 FH extraction challenge, with only the top system convincingly outperforming our approach in terms of precision.

(JMIR Med Inform 2021;9(4):e24020) doi:10.2196/24020

KEYWORDS

information extraction; natural language processing; clinical natural language processing; named entity recognition; sequence tagging; neural language modeling; data augmentation

Introduction

Motivation and Contributions

The widespread use of electronic health records (EHRs) is believed to be one of the key enabling factors leading to the improvement of patient outcomes through data analytics. The analysis of EHRs has been successfully carried out for more than a decade in various health care scenarios [1,2]. Nonetheless, a significant proportion of the information stored in digital patient files is trapped in free-text representations. In particular, family history (FH) reports, vital in the diagnosis and treatment of genetic disorders and familial diseases, such as cardiovascular diseases and cancers, are often stored within EHRs as lengthy textual fields.

In the natural language processing (NLP) subfield of artificial intelligence, information extraction (IE) from free text has been studied for decades. However, IE for biomedical and clinical text is one of the most difficult scenarios for 3 main reasons: (1) entities are complex and diverse [3], (2) clinical text is fragmented and contains shorthand terms, and (3) annotated data are scarce.

We describe a system for extracting information contained in FH reports. The aim of our system is to detect family member mentions (family member [FM] type) and detect mentions of diseases (Observation type). It is developed and evaluated within Track 2 of 2019 N2C2/Open Health Natural Language Processing Shared Task [4], subtask 1.

We leverage pretrained biomedical neural language models (LMs) and combine them with rule-based heuristics and coreference resolution to identify diseases (observations) and FMs in clinical notes. Our main contributions are as follows:

- An entity detection system for FH notes with a state-of-the-art named ennity recognition (NER) model for disease mention detection and a set of heuristics for annotation and normalization of FM mentions
- A detailed evaluation of different transfer learning strategies to improve the annotation of diseases
- A discussion of contributions of individual system components in FM mention detection paired with a detailed error analysis
- An analysis of applicability of coreference resolution to the problem of FM annotation

Our experimental evaluation shows that our system performs better than the median for all systems participating in the shared task by a considerable margin. We believe that an architecture such as ours, which uses domain-specific rules where training data are noisy or scarce, has high applicability in the creation of refined training data sets for FH IE.

Background and Previous Work

EHRs - Context

EHRs, the majority of which contain free text, such as clinical notes, discharge summaries, and pathology reports, have led to an improvement in health care quality by electronically documenting patients' medical conditions [5,6]. EHRs are used for various primary and secondary purposes, such as care process modeling, clinical decision support, biomedical research, and epidemiological monitoring of the nation's health. Although NLP and machine learning (ML) applications in clinical text are receiving attention, the progress is limited because of the lack of shared data sets and tools because of privacy and data confidentiality constraints. To overcome these challenges, efforts have been made by shared task organizers, such as the National Natural Language Processing Clinical Challenges (N2C2), to promote clinical NLP research and provide a standard benchmark to evaluate the performance of the proposed systems.

In next subsections, we introduce some of the related IE techniques and provide a summary of past studies on FH extraction.

Clinical IE

IE is the process of translating free text into structured data. It often includes 2 tasks: (1) NER, where mentions of named entities are identified in free text, and (2) relations between these named entities are identified. In the clinical setting, these entities can be symptoms, drugs, or diseases [6].

Earlier IE systems often relied on expert rules to identify mentions of predefined entities. Rule-based toolkits specialized for clinical text, such as MetaMap [7,8], rely on external knowledge sources of biomedical terms, such as the SPECIALIST lexicon, and use complex rules to identify all possible mention variants of an entity, including acronyms, abbreviations, synonyms, or derivational variants. These tools can usually achieve high precision (when the identified mentions are indeed correct) at the expense of low recall (when many mentions are missed). Another shortcoming of rule-based systems is that expensive human efforts are required to maintain the resources and to expand the rules, enabling them to stay up to date with evolving language use and domain knowledge.

ML-based systems [9,10] replace *hard* rules as *soft* features and estimate the importance (weights) of features using annotated training data. Despite the successful applications of ML-based IE systems, they still display domain discrepancies. That is, the distribution of training data, based on which feature templates are designed and weights are estimated, differs from the data distribution where the system is employed. Therefore, the quality of manually designed feature templates is critical for the system. These features should be informative and should generalize for unseen data.

To alleviate the burden of manually building feature templates, deep learning models have been increasingly applied on clinical IE tasks. A key idea that enables the success of recent deep learning–based models in NLP is that word meanings can be encoded in dense vectors via pretraining on raw text [11-13]. Efforts along this direction in clinical NLP focus on obtaining better word representations for clinical text [14]. For example, Alsentzer et al [15] and Huang et al [16] pretrained Bidirectional Encoder Representations from Transformers (BERT) models on clinical notes and achieved better performance than BERTs pretrained on generic-domain text. Zhang et al [17] investigated strategies to adapt generic-domain embeddings to the clinical domain. Another direction in clinical IE is to identify complex

entities that are less common in generic NLP. For example, Wang and Lu [18] and Dai et al [3] proposed models to recognize overlapping or discontinuous entities that usually represent compositional concepts that differ from concepts represented by individual components.

Previous Work

FH plays an important role in the decision-making process of diagnosis and treatment of medical conditions, as it captures shared genetic variations among FMs. Information such as age, gender, and the degree of relatives are also considered in the risk assignment of various common diseases [19]. Many care process models use FH information for decision making in diagnosis and treatment [5]. Modern health care systems usually record FH through structured forms, including free-text sections, which are filled either by a patient or by a clinician. Polubriaginof et al [20] assessed the quality of the FH captured in EHRs. They found that free-text observations were more comprehensive than structured observations, which motivated our study.

The task of extracting FH from clinical notes is challenging because the information can be spread in the patient's progress notes [21]. In addition, FH information is expressed via relations between named entities and may contain contextual information such as certainty and negation, vital statistics, and age modifiers [22]. If we predict that a patient is at an increased risk of developing a certain disease based on FH, we could potentially diagnose it early, leading to early treatment. Computer-based tools can facilitate the effective use of FH and, therefore, provide better personalized care [20]. To provide comprehensive patient-provided FH data to physicians, there is a need for NLP systems that can extract FH from text. The task of FH IE generally includes NER or relation extraction [23].

Friedlin and McDonald [24] developed a rule-based system, a Regenstrief Data Extraction (REX) tool, for extracting and coding FH data from hospital admission notes. The REX tool first locates and extracts the FH section from the admission notes. It then attempts to identify diseases. This system led to a sensitivity of 93% and a positive predictive value of 97% on the 1 years' worth of hospital admission notes. However, the study was limited to only 12 diseases. Goryachev et al [25] developed a rule-based system to identify and extract FH from discharge summaries and outpatient clinical notes. The Health Information Text Extraction [26], which is built on top of a General Architecture for Text Engineering [27] framework, is used to parse discharge summaries and patient notes. Experiments on a set of 2000 reports yielded 85% precision and 87% recall. The architecture yielded promising results; however, the validation set used in the study was small.

Lewis et al [21] followed a 2-step method that selects candidate FH sentences based on the presence of words such as *mother* or *brother* and then uses a set of dependency-based syntactic patterns to extract appropriate diagnoses and identify the FMs referred to. This study restricted observations to concepts that could be mapped to the International Classification of Diseases, ninth edition, codes. They also limited their work to per-sentence IE without considering any cross-sentence anaphoric or coreference resolution. In our study, we experimented with a coreference resolution and evaluated it in our setup.

Almeida and Matos [28] developed rule-based methods using dependency parsing and a phrase-characteristic extraction approach to extract FH information from clinical notes. They used Stanford CoreNLP [29] to process the data, perform dependency parsing and coreference resolution, and then annotate their data for all FMs and observations. This way, context from previous sentences was also considered. On the N2C2 2019 shared task, which is the same data set that we used, they reported F1 scores of 72% and 74% for the discovery of FMs and observations, respectively. Their approach relied on heuristics to detect arguments of relations, such as using a predefined list of family relationships and diseases or making use of it as arguments in the noun phrases that are detected close to the suspected relationship markers. However, finding relation arguments is challenging because of their variable lengths.

When NER and relation extraction are applied in a pipeline, the error propagates from the NER module to the relation extraction module. To avoid this error propagation, Shi et al [23] proposed a joint learning method that tackles both of these tasks by sharing parameters in a unified neural network framework. The FH IE is performed at different levels, including FMs, observation, and living status and the side of the family (maternal, paternal, and not available). Each input token is represented by word embeddings and corresponding Part-of-Speech embeddings and is given as an input to the bidirectional long short-term memory (BiLSTM) model. Their proposed model ranked first in the 2018 N2C2 FH extraction challenge. They achieved an F1 score of 89% for entity identification and 64% for FH extraction. On the basis of the error analysis, the authors found that a large number of errors are caused by indirect relatives, which can be improved by considering relations among relatives, a feature we incorporate in this study.

Dai [30] formulated the FH IE task as a sequence-labeling problem in which a neural sequence-labeling model was employed along with different tag schemes to distinguish FMs and their observations. They proposed a BiLSTM-Conditional Random Fields (CRFs) model with 3 layers: the character sequence representation layer, the word sequence representation layer, and the inference layer. The proposed method achieved an F1 score of approximately 85% on the test set, which ranked second in the FH entity recognition subtask of the 2018 N2C2 FH extraction challenge. Although the proposed BiLSTM-CRF network is effective in modeling contextual information and label dependencies, it has limitations in that the network can only exploit contexts within individual sequences (sentences) but cannot obtain context from cross-sentence information. To overcome this limitation, Dai et al [31] introduced a neural attention model to exploit cross-sentence information to identify mentions.

Zhan et al [32] fine-tuned the BERT model by including an additional biaffine classifier adapted from the dependency parsing to extract FH mentions.

FH Extraction Task

FH IE, as defined in the N2C2 FH 2019 shared task, includes the following 2 subtasks:

1. Entity identification, including FMs, the side of family (paternal, maternal, and not applicable [NA]), and observation (disease)
2. Relations between FMs, including observations (negated or not) and their living status.

The possible FMs in this task are father, mother, parent, sister, brother, daughter, son, child, grandmother, grandfather, grandparent, cousin, sibling, aunt, and uncle. Other relatives, such as spouses (not blood related), nieces, and nephews were excluded. For first-degree relatives—parents, children, and siblings—the side of family is NA.

In relation extraction, a *living status score* is defined per extracted FM to encode whether they are alive and healthy. In this study, however, we focused only on the entity identification subtask.

Data Set

The data set for the FH task was curated from synthetic English patient notes, which were randomly sampled from the Mayo Employee and Community Health cohort. It contains 216 notes, which we refer to as *documents*, from which 99 documents are for training and 117 for testing. A total of 2 annotators and 1 adjudicator annotated the corpus, with an interannotator agreement of 0.84 for entities and 0.70 for relations. The overall statistics of the corpus are shown in Table 1.

Table 1. Statistics (counts) of entities and relations in the National Natural Language Processing Clinical Challenges family history data set.

Data set's artifact	Training size, n	Test size, n
Document	99	117
Family member	803	760
Observations	978	1062

Evaluation Metrics

For entity extraction, a system extracts either a triplet *(document identifier, family member,* and *side of family)* for FM mentions or a pair *(document identifier* and *text of observation)* for observation mentions. These triplets and pairs were matched against the gold standard.

Observation partial matches are acceptable. For example, *diabetes* is accepted for *diabetes type 2*. The standard F1 score, precision, and recall metrics are used to evaluate the effectiveness of the proposed models as follows:

TP denotes true positive, FP denotes false positive, and FN denotes false negative.

Importantly, recall and precision are defined on sets of annotations pertinent to each document. That is, a document can mention *cancer* multiple times, but detection of any of these mentions contributes to the TP count only once. Conversely, the lack of detection of any of these mentions contributes only once to the FN count.

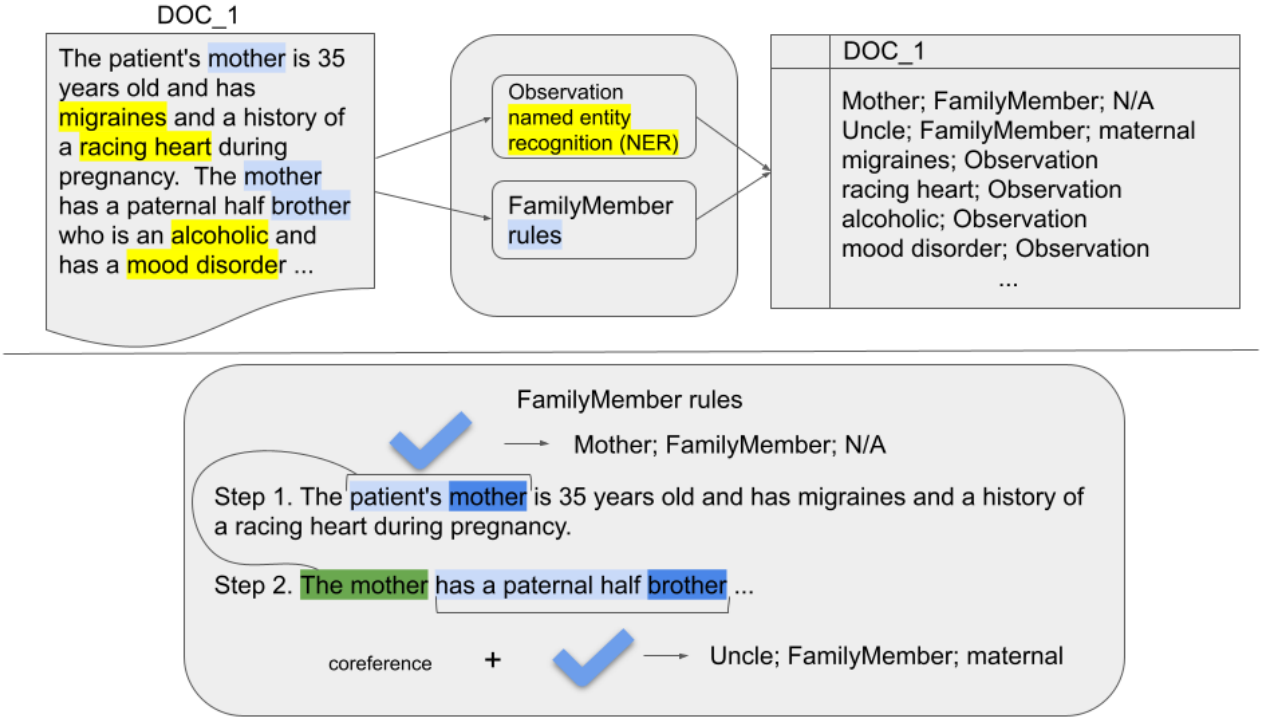
For statistical significance testing, we use a paired approximate randomization test [33] for pairwise comparisons between system variants. We obtain the significance levels by running 9999 pseudorandomized shuffles of the test set results.

Methods

Overview

Our task consists of detecting 2 types of mentions: Observation and FamilyMember. The dual objective of the task is reflected in the design of our system, in which the disease mentions are detected with an ML-based NER component, whereas FamilyMember mentions are detected with a hybrid (rule based, with some ML components) module. The overall architecture, together with the inputs and outputs, is illustrated in the top part of Figure 1.

Figure 1. Overview of the system and the FamilyMember mention detection. N/A: not applicable; NER: named entity recognition.



Observation-NER

Problem formulation

We formulated the observation-NER as a sentence-level sequence tagging problem, in which each word in the sentence is assigned a tag. The tag, which uses the Beginning-Inside-Outside schema [34], can be used to infer whether the word is the first word within a mention or inside a mention or does not belong to any mention.

The sequence tagger we use is a state-of-the-art model: the BERT-CRF model [11,35]. It takes advantage of large-scale pretrained LMs using BERT to create contextual representations for each word and a probabilistic graphical model using CRFs [36] to capture dependencies between neighboring tags.

BERT-Based Encoder

Given a sentence, the tokenizer, coupled with the pretrained BERT model, first converts each word in the sentence into word pieces. That is, if the original word does not exist in the vocabulary of the tokenizer, it will be segmented into several units from the vocabulary [37]. Then, the word pieces are mapped to dense vectors via a lookup table (also known as token embeddings). Finally, the sum of token embeddings and positional embeddings, which indicate the position of each token in the sequence, is fed into a stack of multihead self-attention and fully connected feed-forward layers [38]. Following the work by Devlin et al [39], we use the final outputs corresponding to the first word piece within each word as the word representation.

CRFs in NER

Instead of assigning a tag to each word independently, we modeled them jointly using CRFs [40]. That is, given a sequence of word representations $X = (x_1, x_2,..., x_n)$, we aim to predict a sequence of tags ($Y = (y_1, y_2,..., y_n)$) that has the maximum probability over all possible tag sequences. This conditional probability can be calculated using the following equations:

$A_{i,j}$ is the compatibility score of a transition from the tag i to tag j and $P_{i,j}$ is the score of the tag j given X_i.

The parameters from both BERT and CRFs are trained to maximize the conditional probability of the gold tag sequence, given the training sentences.

Enhancing BERT

The vanilla BERT model is pretrained using generic-domain data such as books and Wikipedia, which are very dissimilar to task data. A previous study has shown that the effectiveness of pretrained LMs is highly affected by the similarity between source pretraining and target task data [41].

Thus, we explored 2 approaches to improve the effectiveness of vanilla BERT on the target task: domain-adaptive pretraining (DAPT) [42] and intermediate-task pretraining (ITPT) [43].

DAPT Approach

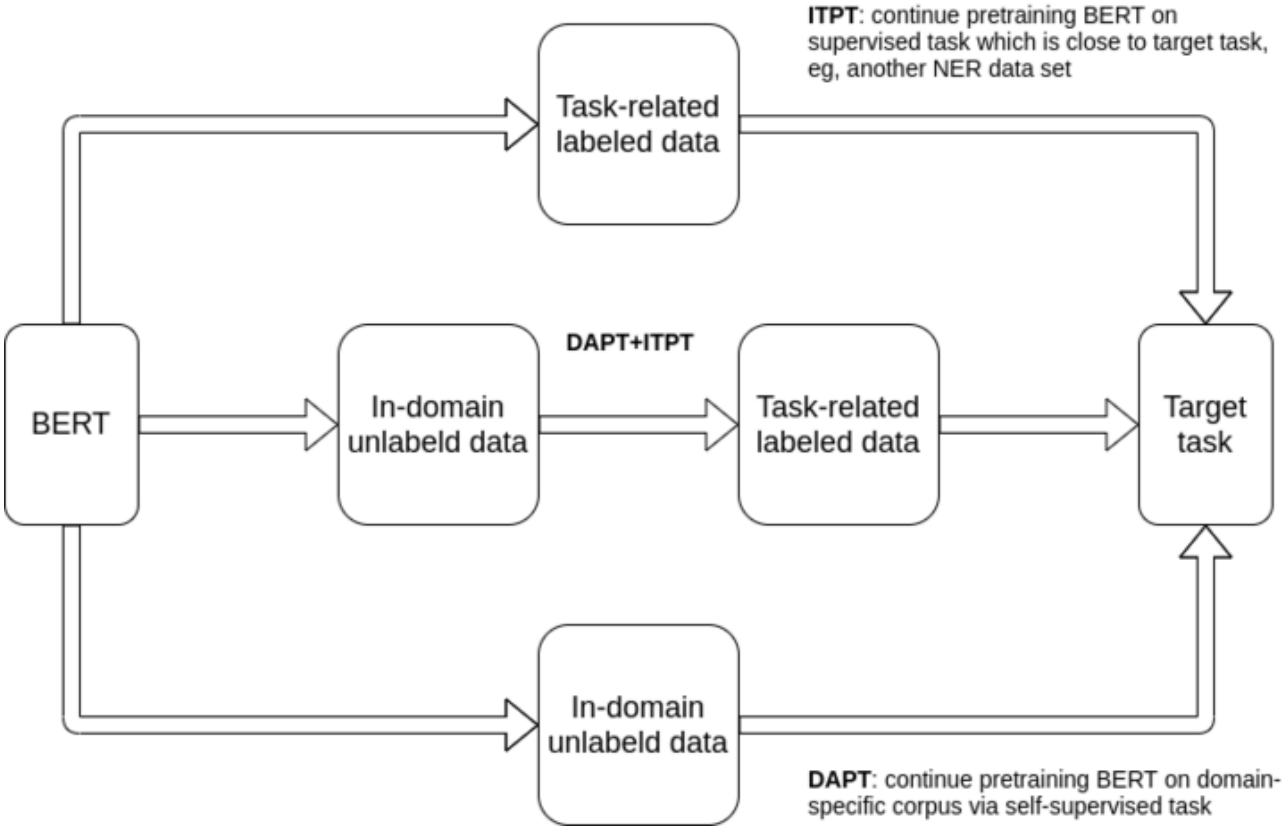
The DAPT approach consists of continued pretraining of BERT on a large volume of unlabeled in-domain text. The training uses a masked language modeling objective to adapt the weights of BERT to the domain of the target task. We use BioBERT [44] and ClinicalBERT [15] as proxies for DAPT. These models employed continued BERT pretraining on biomedical scientific papers and hospital discharge summaries.

ITPT Approach

ITPT consists of the pretraining of BERT together with CRFs by training on a target task-related NER data set (usually annotated with similar entity types). The training uses the sequence tagging objective to jointly optimize the weights of both BERT and CRF layers toward the specific task. We used the National Center for Biotechnology Information (NCBI)-disease [11] data set for ITPT. This data set consists of 793 PubMed abstracts that are fully annotated at the mention and concept levels. It contains 6892 disease mentions, which are mapped to 790 unique disease concepts. The motivation for this choice is 2-fold. First, we used the NCBI-disease data set because of the semantic overlap between the *Disease* and *Observation* concepts and because of the size of the NCBI-disease data set, which is larger than the in-domain data (NCBI-disease consists of nearly 800 documents, with almost 7000 disease mentions). Second, this choice results in a more direct comparison with our off-the-shelf baseline, which was trained on the same NCBI-disease corpus (our experimental setup is explained in more detail at the beginning of the *Results* section).

We also explored the combination of these 2 approaches, that is, DAPT and ITPT. A high-level comparison of these 3 approaches is presented in Figure 2.

Figure 2. Different approaches to enhance Bidirectional Encoder Representations from Transformers for a given target task (domain-adaptive pretraining and intermediate-task pretraining). BERT: Bidirectional Encoder Representations from Transformers; DAPT: domain-adaptive pretraining; ITPT: intermediate-task pretraining; NER: named entity recognition.



After DAPT or ITPT, we continued to fine-tune the model weights of the target task's training data.

Owing to the aforementioned semantic overlap between classes of interest, NCBI-disease was our first choice for ITPT. Nonetheless, for the sake of completeness, we also present an ITPT evaluation (for all DAPT configurations) for other publicly available candidate data sets, which involve annotation of diseases, that is, Integrating Biology and the Bedside (i2B2) 2010 [45] and Shared Annotated Resources - Conference and Labs of the Evaluation Forum 2013 [46].

Details of implementation and training of our BERT-CRF models are outlined in Multimedia Appendix 1.

FamilyMember Mentions

The FamilyMember mentions' detection often requires an out-of-sentence context to make correct inferences. In the example shown in Figure 1, an out-of-sentence context is needed to resolve coreference to "the mother" and correctly normalize the *brother* or *uncle* mention. Another instance where a broader document context is required is deciding whether the information provided in a given fragment of an FH note pertains to the patient's or their partner's family. The task is focused on extracting the information on the patient's blood relatives; therefore, the mentions of the partner's family should not be annotated, although at the sentence level, the information can be identical.

Given the moderate size of the training corpus of approximately 100 documents and the complexity of the FamilyMember normalization task, which entails multiple entity classes, during our participation in the shared task, we opted for a rule-based approach enhanced with some ML elements. This early design choice determines the scope of our focus; however, we compare

our approach with state-of-the-art deep learning baselines trained on the available in-domain data.

In our hybrid system, the documents are analyzed sentence by sentence with a series of pattern-matching rules. The previous sentence is used as context when producing FamilyMember annotations for a given sentence (we split each document into sentence-level bigrams). We experimented with a state-of-the-art coreference resolution model *neuralcoref* [47]. Coreference resolution is used on pairs of consecutive sentences to incorporate context information from possessive pronouns (eg, *her son*) or other third-person pronouns (*she has a son*) and alternative references (eg, *this woman has a son*).

To detect paragraphs of the notes containing information on the partner's family, we incorporate a state-of-the-art text classification model. Owing to the lack of dedicated partner-paragraph annotations, we fine-tune a BERT [39] model on the available training data. We formulate the task as a binary classification problem, where the model predicts whether a given paragraph is valid (containing patient-focused information). The previous paragraph is also fed to the model to provide contextual information. We derive validity from the existing training data. That is, a paragraph is valid if at least one annotation is present within its scope in the training data set. At the annotation time, we skip sentences predicted to be part of invalid paragraphs by the model.

Our first step is to predict the validity of each of the paragraphs of an FH note using a BERT-based paragraph filter. This step results in filtering out the paragraphs that are predicted to be invalid by the model. We then iterate all the remaining document sentences to create sentence-level annotations. These annotations are then put into one document-level annotation set. The procedure for FamilyMember mentions' detection within the scope of a sentence, given a previous sentence and its annotations as context for coreference, consists of the following steps:

1. Check if a sentence is not part of an invalid paragraph. If it is, we skip to the next sentence.
2. Detect candidate mentions in the second sentence (predictions for FamilyMember for the previous sentence are already available); candidate mentions are occurrences of words denoting family relationships relevant to the task as per the task definition, such as *brother, sister, mother, and father.*
3. Build a graph of candidate-candidate relationships. For example, an expression *mother's sister* would result in vertices *mother* and *sister* and a directed edge from *mother* to *sister*; this graph incorporates coreference information.
4. Generate FamilyMember annotations from the graph structure according to a set of rules. For example, the mother-to-sister structure would generate annotations Mother-NA (not applicable) and Aunt-Maternal.

To build a graph of candidate-candidate relationships, we look for specific linguistic patterns between pairs of adjacent candidate mentions. These patterns are "X's *Y, X*has/had *Y, Y of *X," where X and Y denote candidates such as *brother*, *sister*, or *uncle* and the * symbol denotes a wildcard matching any text. We also detect candidate-candidate relationships as adjacency of candidate mentions to expressions linked with coreference resolution to other candidates or FamilyMember mentions from the previous sentence. For example, if in the sentence pair "Mr. Williams' mother is alive and well. She has an older sister...," the word *mother* is annotated with as *Mother* and word *She* falls into the same coreference cluster as the word *mother*, and then a Mother-Sister relationship will be added to the graph as an X*has*Y pattern is triggered. Downstream, this relationship is used to normalize the annotation of *the sister* according to the rules (to Aunt, Maternal).

To convert the candidate graph to a final representation, we apply a set of rules to each of the vertices. The procedure, together with these rules, is presented in the pseudocode in Figure 3.

In addition, we apply a simple heuristic approach to determine the family side for those annotations where the side of the family cannot be determined by inspecting the parent node in the candidate graph. We look for the last occurrence in the text of words *maternal*, *mother* or *paternal*, *father*, before the given candidate is mentioned. The maternal or paternal status is determined according to this last occurrence. We only assign *NA* if none of these words appear in the document before the candidate is mentioned.

Figure 3. A pseudocode representation of the rule-based processing.

```
method convert_to_annotations(graph G):
  annotations=Set()
  for candidate in G.vertices:
    parent_candidate = G.get_parent_vertex(candidate)
    if parent_candidate == NULL:
      annotations.add(candidate)
    else:
      // R1: Uncle Rule
      if (parent_candidate.relationship in ['mother', 'father']
        AND candidate.relationship=='brother'):
          side = determine(parent_candidate)
          annotations.add(FamilyMember('uncle', side))
      // R2: Aunt Rule
      if (parent_candidate.relationship in ['mother', 'father']
        AND candidate.relationship=='sister'):
          side = determine(parent_candidate)
          annotations.add(FamilyMemeber('aunt', side))
      // R3: Grandparents Rule
      if (parent_candidate.relationship in ['mother', 'father']
        AND candidate.relationship in ['mother', 'father']):
          side = determine(parent_candidate)
          annotations.add(FamilyMember('grand'
                                       +candidate.relationship, side))
      // R4: Sibling's Kids Rule
      if (parent_candidate.relationship in ['brother','sister','sibling']
        AND candidate.relationship in ['son', 'daughter', 'child']):
          do_nothing()
      // R5: Cousin Rule 1
      if (parent_candidate.relationship in ['aunt', 'uncle']
        AND candidate.relationship in ['son', 'daughter', 'child']):
          side = determine(parent_candidate)
          annotations.add(FamilyMember('cousin', side))
      // R6: Cousin Rule 2
      if (NOT match(parent_candidate.relationship, 'grand*')
        AND candidate.relationship=='cousin'):
          side = determine(parent_candidate)
          annotations.add(FamilyMember('cousin', side))
```

Results

Observation Extraction

The gold standard tags are recreated naively by string matching the gold annotations provided. For example, given a gold annotation *mental retardation*, we find all occurrences of this annotation in the corresponding document and assign *B-Observation I-Observation* tags to all the identified spans. We select the first 18 documents from the training set as the development set. The trained model that is most effective on the development set, measured using the span-level F1 score, is used to evaluate the test set. In addition to the different variants of BERT models, we use an off-the-shelf disorder NER model as the baseline [48].

We present the results of our main experiments with the Observation annotation in Table 2. We achieve the best results for a BERT model using both DAPT and ITPT. We provide a detailed discussion of these results in the *Discussion* section.

The results of the additional ITPT experiments with i2b2 2010 and ShARe-CLEF (Shared Annotated Resources-Conference and Labs of the Evaluation Forum) 2013 are presented in Table 3. The results indicate that, although improvements from ITPT alone are comparable with those obtained with the NCBI-disease data set, the DAPT+ITPT combination with the alternative disease annotation data sets is less successful.

Table 2. Evaluation results on Observation concepts in the test set[a].

Method	Precision		Recall		F1 score	
	Value	*P* value	Value	*P* value	Value	*P* value
Stanza [48]	81.5	N/A[b]	75.0	N/A	78.1	N/A
BERT[c] (baseline), mean (SD)	70.7 (2.7)	N/A	87.3 (1.5)	N/A	78.1 (1.1)	N/A
DAPT[d] (BioBERT), mean (SD)	73.4[e] (2.2)	<.001	86.5 (1.7)	<.001	79.4 (0.6)	.08
DAPT (ClinicalBERT), mean (SD)	76.2[e] (3.5)	<.001	83.4 (3.1)	<.001	79.5[e] (1.0)	.002
ITPT[f] (NCBI[g]-disease), mean (SD)	75.0[e] (1.8)	<.001	85.3 (1.1)	>.99	79.8[e] (0.6)	<.001
DAPT (BioBERT)+ITPT (NCBI-disease), mean (SD)	77.7[e] (2.6)	<.001	85.1 (2.8)	.08	81.1[e] (1.1)	<.001
DAPT (ClinicalBERT)+ITPT (NCBI-disease), mean (SD)	78.6[e] (3.2)	<.001	84.4 (1.5)	.56	81.3[e] (1.2)	<.001

[a]Document-level precision, recall, and F1 score are reported using official evaluation scripts.

[b]N/A: not applicable.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[c]DAPT: domain-adaptive pretraining.

[e]Represents results that are significantly better than the Bidirectional Encoder Representations from Transformers baseline (approximate randomization test; *P*=.05). Although the recall of baseline Bidirectional Encoder Representations from Transformers is the highest, the differences are not significant except those for 2 domain-adaptive pretraining variants.

[f]ITPT: intermediate-task pretraining.

[g]NCBI: National Center for Biotechnology Information.

Table 3. Evaluation results on Observation concepts in the test set for different intermediate-task pretraining and domain-adaptive pretraining combinations[a].

ITPT[b] and BERT[c] model	Precision, mean (SD)	Recall, mean (SD)	F1 score, mean (SD)
BERT	75.0 (1.8)	85.3 (1.1)	79.8 (0.6)
NCBI[d]-disease			
+DAPT[e] (BioBERT)	77.7 (2.6)	85.1 (2.8)	81.1 (1.1)
+DAPT (ClinicalBERT)	78.6 (3.2)	84.4 (1.5)	81.3 (1.2)
BERT	71.6 (3.4)	88.9 (2.4)	79.2 (1.5)
i2b2[f] 2010			
+DAPT (BioBERT)	75.6 (1.9)	86.2 (1.4)	80.5 (1.4)
+DAPT (ClinicalBERT)	73.2 (2.0)	89.0 (1.8)	80.3 (0.7)
BERT	70.7 (2.7)	88.7 (1.5)	78.6 (1.3)
ShARe-CLEF[g] 2013			
+DAPT (BioBERT)	72.9 (2.5)	88.3 (2.3)	79.8 (0.8)
+DAPT (ClinicalBERT)	74.2 (2.6)	86.5 (3.8)	79.8 (0.9)

[a]Document-level precision, recall, and F1 score are reported using official evaluation scripts.

[b]ITPT: intermediate-task pretraining.

[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]NCBI: National Center for Biotechnology Information.

[e]DAPT: domain-adaptive pretraining.

[f]i2b2: Integrating Biology and the Bedside.

[g]ShARe-CLEF: Shared Annotated Resources-Conference and Labs of the Evaluation Forum.

FamilyMember Extraction

We experimented with the different settings of our approach by evaluating, both on training and test sets, different combinations of the elements of our systems. In addition, we experimented with removing *child*, *sibling*, *parent*, and *grandparent* from the set of relationships, as we hypothesized that the corresponding words are not often used to introduce a particular FM (eg, “She

has 4 siblings: two brothers and two sisters"). We obtained the best results on the test set for a system with a restricted set of relationships, using all the rule 1 (R1) to R6 and with BERT-based paragraph filtering, but without the coreference resolution.

The performance of the best system is presented in the first row of Table 4. Subsequent rows demonstrate the impact of modifying the best run by adding the remaining relations (row 2), adding coreference resolution (row 3), removing the BERT paragraph filter (row 4), and removing rules R1-R6 (rows 5-10). Row 11 shows a baseline system with no rules, no paragraph filter, and no coreference resolution, working with the full set of relations.

We compare the results obtained with our hybrid approach with those obtained with a BERT-CRF baseline, identical to those employed for disease annotation. For the sake of completeness, we include baseline results for domain-adapted flavors of BERT—BioBERT and ClinicalBERT.

Table 4. FamilyMember detection for different settings of the system.

Row	Number of relations[a]	Coreference	R1[b]	R2	R3	R4	R5	R6	BPF[c]	Training, precision (*P* value)	Test, precision (*P* value)	Training, recall (*P* value)	Test, recall (*P* value)	Training, F1 score (*P* value)	Test, F1 score (*P* value)
(1)	11	—[d]	✓[e]	✓	✓	✓	✓	✓	✓	90.34[f]	81.38[f]	85.60[f]	82.91	87.91[f]	*82.14*[f,g]
(2)	15	—	✓	✓	✓	✓	✓	✓	✓	86.00[f,h] (<.001)	73.73[f,h] (<.001)	89.35[f,h] (<.001)	*86.67*[g,h] (<.001)	87.64[f] (.68)	79.68[f,h] (<.001)
(3)	11	✓	✓	✓	✓	✓	✓	✓	✓	88.07[f,h] (<.001)	77.59[f,h] (<.001)	83.05[f,h] (<.001)	79.78[f,h] (<.001)	85.49[h] (<.001)	78.67[f,h] (<.001)
(4)	11	—	✓	✓	✓	✓	✓	✓	—	87.42[f,h] (<.001)	77.98[f,h] (<.001)	86.50[f,h] (.03)	83.85 (.13)	86.96[f,h] (.04)	80.81[f,h] (.01)
(5)	11	—	—	✓	✓	✓	✓	✓	✓	90.04[f] (.51)	81.61[f] (.40)	85.45[f] (>.99)	82.13 (.07)	87.69[f] (.51)	81.87[f] (.35)
(6)	11	—	✓	—	✓	✓	✓	✓	✓	90.06[f] (.51)	*81.71*[f,g] (.14)	85.60[f] (>.99)	81.97[h] (.03)	87.77[f] (.51)	81.84[f] (.19)
(7)	11	—	✓	✓	—	✓	✓	✓	✓	*90.64*[f,g] (.50)	81.60[f] (.59)	85.75[f] (>.99)	81.34[f] (.09)	*88.13*[f,g] (.50)	81.47[f] (.25)
(8)	11	—	✓	✓	✓	—	✓	✓	✓	89.79[f] (.14)	80.15[f,h] (.004)	85.75[f] (>.99)	83.54 (.13)	87.73[f] (.26)	81.81[f] (.27)
(9)	11	—	✓	✓	✓	—	—	✓	✓	90.62[f] (.73)	81.13[f] (.56)	85.45[f] (>.99)	82.91 (>.99)	87.96[f] (>.99)	82.01[f] (.74)
(10)	11	—	✓	✓	✓	✓	✓	—	✓	90.34[f] (>.99)	81.38[f] (>.99)	85.60[f] (>.99)	82.91 (>.99)	87.91[f] (>.99)	*82.14*[f,g] (>.99)
(11)	15	—	—	—	—	—	—	—	—	81.52[h] (>.001)	69.01[h] (>.001)	*89.95*[h] (>.001)	84.48 (.38)	85.53[h] (.01)	75.96[h] (>.001)
(12)[i]		—	—	—	—	—	—	—	—	N/A[j]	79.72[f] (.33)	N/A	81.35 (.45)	N/A	80.35[f] (.26)
(13)[k]		—	—	—	—	—	—	—	—	N/A	81.55[f] (.95)	N/A	81.03 (.462)	N/A	81.29[f] (.70)
(14)[l]		—	—	—	—	—	—	—	—	N/A	82.71[f] (.62)	N/A	79.47 (.17)	N/A	81.06[f] (.62)

[a]Denotes size of the relationship set.

[b]R1-R6 denote uncle rule, aunt rule, grandparents rule, sibling's kids rule, cousin rule 1, and cousin rule 2, respectively.

[c]BPF: Bidirectional Encoder Representations from Transformers–based paragraph filter.

[d]Not available.

[e]Denotes that the corresponding rule is applicable.

[f]Denotes statistically significant ($P \leq .05$) difference from the baseline (row 11).

[g]We report the *P* values corresponding to the test against the best system. Highest measured value is denoted in italics.

[h]Denotes statistically significant ($P \leq .05$) difference from the top system (row 1).

[i]Bidirectional Encoder Representations from Transformer-Conditional Random Field baseline results on the test set for Bidirectional Encoder Representations from Transformer.

[j]N/A: not applicable.

[k]Bidirectional Encoder Representations from Transformer-Conditional Random Field baseline results on the test set for BioBERT.

[l]Bidirectional Encoder Representations from Transformer-Conditional Random Field baseline results on the test set for ClinicalBERT.

Discussion

Principal Findings

Challenges of recognizing diseases in clinical narratives, such as a wide variety of naming patterns and data anonymization, have been widely studied in the literature [49,50]. Therefore, we provide only a discussion on disease identification that relates specifically to FH extraction tasks and a detailed discussion on FM identification.

Annotation of Observations

Impact of Domain Adaptation

From Table 2, we observe that both DAPT and ITPT can improve over the baseline of the BERT-CRF model, and combining these 2 approaches first with DAPT and then ITPT achieves the best F1 score. On the basis of this result, we argue that DAPT and ITPT can complement each other. In other words, they enhance pretrained LMs by providing different inductive biases. We hypothesize that in the ideal scenario, DAPT enforces the model to be more compatible with the language distribution of the target data and ITPT enforces the model to pay more attention to features that are informative to the NER task.

The aforementioned hypothesis can also be used to explain the results presented in Table 3. We observe that NER *problem* and *disorder* classes (i2b2 and ShARe-CLEF, respectively) are less semantically aligned to our *Observation* class than *Disease* from NCBI. In particular, a large proportion of the *problem* and *disorder* mentions could be classified as *symptoms* (eg, *headaches*, *fever*, and *pain*). Disease names annotated in NCBI-disease seem closer to our target task's *Observation* entity category. It is possible that the alternative ITPT data sets provide an isolated improvement by exposing the model to documents similar to that of the target corpus but offer little improvement when combined with DAPT (which, we assume, already provides this inductive bias).

Error Analysis

To provide some insight into the role of task-specific fine-tuning with BERT-like models, we provide a detailed error analysis performed on the outputs generated by an off-the-shelf baseline (trained on the NCBI-disease data set, not tuned on the FH extraction task) and our best system, which is ClinicalBERT with ITPT on the NCBI-disease data set.

The error analysis, apart from counts of FP and FN errors, involves a fine-grained classification of 50 errors of each type (FN/FP) per model. The errors were sampled by taking the first 50 errors of a given type from the output log with randomly shuffled documents.

We classify FP errors into the following categories:

- True FPs: The span does not cover a valid Observation candidate. For example, “The patient's father had six-vessel bypass surgery at age 56.”
- Relative error: The Observation mention by itself is identified correctly but is linked to a relative who is not a suitable candidate for a FamilyMember annotation (eg, a great-uncle would be an example of a too distant relationship, according to the annotation guidelines provided for the task). Importantly, this class of FP errors also covers disease mentions pertaining to the family of the patient’s partner (thus, not related by blood); for example, “His [husband's] brother died at age 14 of suicide and was thought to have depression.” Note that errors are classified as relative errors if the identified Observation looks correct, and it can be linked to a non-FamilyMember; the annotation is missing from the gold standard test data set; for example, the gold standard annotations expect a span containing stomach cancer, a string that does not appear in the corresponding document.
- Nonobservation errors: FNs where the gold-standard annotation is missed by the system, although it appears in the document, but it could be debated whether it constitutes an actual Observation. For example, “She has some freckles.”
- Questionable and nonerrors: The candidate mention looks correct and is linked to a valid FamilyMember candidate. For example, “Mrs. William's sister has had three miscarriages and a son.”
- Trauma or procedure errors: The predicted span includes a name of a procedure or a traumatic injury. For example, “These last two maternal aunts have had hysterectomies.”
- Negation errors: The predicted span covers a valid Observation candidate, but it appears in a negated and often general context. For example, “There is no known history of ADHD or schizophrenia.”

We propose the following categorization for the FN errors:

- True FNs: An actual valid observation was missed by the model. For example, “Her father is 53 with high cholesterol.”
- Gold standard errors: Errors in the gold standard.
- Mental health/substance abuse–related errors: FNs where the models fail to annotate mental health conditions or addictions. We present this special case of true FNs separately, as the evaluated models particularly struggle with detection of this type of observations. For example, “Maternal grandfather, age 67, smokes but is healthy.”

Overall, the off-the-shelf baseline yielded 166 FPs and 244 FNs, with 733 correct annotations. ClinicalBERT-ITPT generated 150 FPs, 172 FNs, and 805 correctly identified mentions. For Stanza, the off-the-shelf baseline values are shown in Table 2. For ClinicalBERT+ITPT, we analyze the run that achieves the highest F1 score among all 5 experimental runs (0.8333 F1 score, 0.8429 precision, and 0.8240 recall).

Table 5 shows the distribution of the error classes over the evaluated sample of FPs. The distribution of the FN errors is shown in Table 6.

An inspection of FPs reveals that for both models, the main source of error is the annotation of observations pertaining to FMs that are not related by blood to the patient (eg, partner's family) or the family relation is too distant (eg, great-grandparents). The BERT-based model alleviates this problem by fine-tuning, at least to a certain degree. However, as the observation-NER is done on a stand-alone basis (ie, without the joint modeling of Observation and FamilyMember spans), the context awareness of the BERT-based model regarding family relationships remains low.

Both models lead to approximately 20% of FPs that appear to be correct; however, they are not present in the gold standard annotations.

Table 5. Results (counts) of error analysis for false-positive errors for the Observation entity type.

Error type	Stanza [48], n	ClinicalBERT[a] with ITPT[b], n
Relative	31	20
Nonerror	10	9
Trauma or procedure	2	6
Negation	7	9
True	0	6

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]ITPT: intermediate-task pretraining.

Table 6. Results (counts) of error analysis for false negative errors for the Observation entity type.

Error type	Stanza [48], n	ClinicalBERT[a] with ITPT[b], n
Gold standard	4	4
Nonobservation	14	15
Mental or substance	13	6
True	19	25

[a]BERT: Bidirectional Encoder Representations from Transformers.

[b]ITPT: intermediate-task pretraining.

The BERT-based model is more likely to correctly annotate spans of medical procedures or traumatic injuries. This may be a consequence of fine-tuning. Interestingly, these entities are identified inconsistently throughout the data set; that is, examples of this class can be found both in FPs ("These last two maternal aunts have had hysterectomies" where hysterectomies is an FP) and FNs ("The patient's maternal grandmother died at 83 of diabetes and asthma and had a broken hip," where *the broken hip* is an FN, undetected by the system).

Both models produce a similar proportion of errors resulting from annotating negated or general contexts (not pertaining to a specific FM). For both models, spans of this type appear among FPs ("There is no known history of ADHD or schizophrenia," attention-deficit hyperactivity disorder [ADHD] and schizophrenia are the erroneous predictions of the systems) and FNs ("Overall, the family history is not significant for mental retardation, birth defects, multiple miscarriages or neonatal death, or known genetic conditions"; "genetic conditions" is present in the gold standard but missed by the systems).

Finally, the BERT-based model makes some mistakes by selecting spans that do not correspond to valid observations. This might be because of the model being fine-tuned on a small amount of noisy data (examples of negated contexts and traumas/procedures are mentioned earlier).

The analysis of FNs for both models shows a similar trend regarding true errors. That is, a large proportion of the observations missed by both models corresponds to expressions such as "They do not look different from other members of the family, and do not have any major internal birth defects," where the missed span appears in a negative context. Interestingly, this is also true for the off-the-shelf model, which may suggest an inherent problem with negations (as the noisy fine-tuning data cannot be blamed with the off-the-shelf model, which does not undergo fine-tuning altogether).

Similarly, both models struggle with detecting questionable observations. For example, "She attended elementary school and could not walk until the age of 0," where "could not walk" is the gold standard annotation.

A key noticeable difference between the models is that the fine-tuned ClinicalBERT with ITPT does a better job at detecting mental disorders and behavioral traits (eg, "Maternal grandfather, age 67, smokes but is healthy"), resulting in a 50% decrease in this type of errors. This suggests that fine-tuning is beneficial.

Learning Points From Annotation of Observations

The first learning point from our experiments is that an off-the-shelf state-of-the-art model trained for annotating diseases on the NCBI-disease data set provides a strong baseline, which yields much fewer true FP errors than the precision alone suggests. The analysis of the FPs shows that most of the predictions from the models are actually correct in the sense that they correctly identify Observation candidates. It is the detection of only those mentions that are linked only to specific FMs that are the most problematic. This type of constraint is inherently application specific (eg, if the aim were to assess the genetic risk of the child from FH notes collected during pregnancy, then Observations from the patient's partner FH would also be relevant). This means that an off-the-shelf model may be a good starting point for some applications.

Second, we have demonstrated a cumulative value of DAPT and ITPT. This finding highlights an important advantage of the BERT-CRF-based architecture over other state-of-the-art NER approaches, such as BiLSTM. BERT-based architectures offer an out-of-the-box transfer learning framework, with a

focus on sharing models domain adapted on huge corpora (eg, BioBERT and ClinicalBERT used here).

We have also identified the key improvements achieved via fine-tuning by exposure to actual in-domain training data. The fine-tuned model provides better recall, at the expense of precision, achieving the highest aggregated F1 score (although the best-case scenario, ie, the best of our 5 models, outperforms the baseline both in terms of recall and precision). Fine-tuning contributes to better adjustment to the specificity of the task, such as tying Observations to particular FMs. More importantly, however, the model learns to detect behavioral and mental health issues from limited training data, thereby providing a qualitative improvement. This improvement, in this particular evaluation scenario, outweighs the downside of tuning the model on a relatively noisy and low-volume sample of in-domain data (which may result in some loss of precision).

Finally, we observe that, even in the fine-tuned model, there is still room for improvement with respect to adherence to the restrictions relating to the interplay between the observations and FMs. This means that we do not fully capitalize on BERT's capability to capture long-distance relationships in text. In fact, in our experiments, raw BERT-CRF yields a similar F1 score to that of the long short-term memory–based off-the-shelf baseline. We hypothesized that using an alternative training approach or using a network that jointly models both entity types could be the key to better alignment to this specific task.

Annotation of FMs

Impact of Different System Elements

To provide more insight into the effectiveness of our rule-based approach in FM detection, we analyze the errors generated by our best system in row 1 of Table 4.

An ablation study, where we disable one rule at a time, is presented in Table 2. It can be immediately noted that R6 (cousin rule 2; row 10 of Table 4) is a nonfactor. Indeed, it only changes the default behavior (of adding one annotation per surface form, without changing the relationship type), when *cousin* annotation is affected by a candidate mention of any of the grandparents (eg, “grandmother's cousin”). In such cases, the *cousin* mention would not be added to the output. The results show that such an interaction between mentions is not detected in either of the training or test data sets.

In addition, removal of R3 (grandparents rule; row 7 of Table 4) and R5 (cousin rule 1; row 9 of Table 4) minimally increases the effectiveness of our system on the training data set, which does not hold for the evaluation of the test set. Elimination of all other rules (R1, R2, and R4 corresponding to rows 5, 6, and 8, respectively, of Table 4) from the best system impacts the results negatively consistently across data sets.

BERT-based parameter filtering (absent in row 4 of Table 4) impact in the test set evaluation can be seen as a sanity check. As the model was trained on the entire training data set, we assume that it can determine which paragraphs should yield no annotations, as this data set was seen at the training time. Therefore, recall is almost unaffected (we use a cutoff threshold lower than 0.5, which explains the minor change), whereas precision improves, as no annotations are generated from the paragraphs that contain no gold standard annotations. In the test set evaluation, we can see that the BERT-based paragraph filter increases effectiveness in terms of F1 score, but there is some precision-recall trade-off. The increase in precision of approximately 3.5% points comes at a cost of approximately 1% decrease in recall. This decrease in recall can be attributed to 2 types of situations:

- Paragraphs are filtered out correctly but contribute to an annotation missed elsewhere in the text.
- Paragraphs are incorrectly filtered out (misclassified).

By comparing rows 1 and 2 of Table 4, we can see that restriction on the relationship set works in the same direction as BERT-based filtering, which uses only specific relations to increase precision at the expense of recall. Results indicate that it yields a better F1 score; therefore, in terms of F1 optimization, the gain in precision outweighs the loss in recall.

Our experiments with coreference resolution, row 3 of Table 4, show that applying conference resolution within a rule-based system does not improve the annotation effectiveness. Error analysis indicates that coreference resolution often gets it wrong in grammatically ambiguous cases, such as “The patient's mother is 61 and well. Her brother, aged 21, is healthy....” The coreference module, which is trained in isolation from the task, resolves pronouns in a strictly language-focused manner. Resolving the *her* pronoun as a reference to the patient's mother (and consequently producing an *Uncle* annotation) is grammatically plausible but is unlikely from the annotation standpoint. The results and subsequent analysis indicate that the use of coreference leads to an accumulation of such cases, thereby reducing annotation effectiveness.

The results do not point to the standout importance of a single specific technique of those included in our best-performing system. Nonetheless, a combination of the rules with BERT-based filtering and a refined subset of family relationships improves the test F1 score by almost 6% points. We believe that this finding points to the accumulative potential of small improvements in rule-based systems.

Comparisons With the BERT-CRF Baseline

Although our best system (row 1) and other variants yielding similar performance outperform the BERT-CRF baselines (rows 12-14) in absolute values over all metrics, these differences are not statistically significant. Our initial assumption was that the N2C2 FH extraction training data set is too small to successfully train a fully ML-based model for this problem. This assumption ultimately led us to develop a hybrid solution to the FamilyMember annotation problem.

The lack of statistical significance in the advantage of our hybrid model against a strong neural (state-of-the-art) baseline suggests that the assumption was not entirely valid. In fact, the relatively strong performance of the BERT-CRF baseline indicates that this model can cope with the FamilyMember annotation and normalization, despite the small size of the training data set. However, this result also suggests that our hybrid model yields results comparable with those of a state-of-the-art ML model. It is worth noting that a purely rule-based version of our system

(without ML components whatsoever; row 4) still yields comparable results, which would make it an effective, simple baseline (without the need for retraining any of the system's elements) to be considered as a starting point reference for real-world deployment of FH annotation systems. Nevertheless, the impact of individual rules still needs to be considered in the context of the target corpus and task.

Finally, our hybrid system allows for relatively intuitive prioritizing of specific performance aspects (eg, prioritizing recall over precision) by tuning system settings (row 2 of Table 2).

Sentence-Level Error Analysis: FP

We present the findings of a full error analysis performed on the results of our best-performing system (row 1 of Table 2) on the test data. To classify the errors, we examine individual instances in which sentence-level annotations contributed to incorrect predictions. The percentages correspond to sentence-level observations. For example, in a hypothetical passage "Mrs. X has one child, a healthy son. She also has a healthy brother whose partner recently gave birth to a daughter, and a sister, who also gave birth to a healthy daughter," the incorrect annotation *daughter* counts twice (once per occurrence).

We categorize the FPs into the following classes: nonerrors (annotations that we believe to be correct but are not present in the gold standard), nonmentions (relationship word is used but does not denote this particular FM), partner's family (annotations pertinent to the family of the partner rather than the patient), deficiency of rules (when an expression is worded in a way that the rules miss it altogether or produce an undesired output), lack of coreference (context from outside the sentence is missing to produce a correct annotation), wrong family side (maternal/paternal heuristic failing), and other (the annotation looks fine, but even after reading the entire note, we were not able to tell if it is an actual nonerror). We analyze all errors detected on the test set and provide counts for each of the classes together with examples (Table 7).

Table 7. Error classes for false positives with counts and examples obtained for the best-performing FamilyMember extraction.

Class	Count, n	Example	Prediction
Partner's family	39	"Mr William's [from context: Mr Williams is the husband of the patient—Ms Williams] father has a brother who is currently healthy"	(Uncle, paternal)
Nonmention	38	"States on her father's side, 'there is untreated depression'"	(Father, N/A[a])
No coreference	32	"She [sister] has a 2-year-old son"	(Son, N/A)
Nonerror	31	"Mrs Alexander's paternal grandmother reportedly had one miscarriage"	(Grandmother, paternal)
Rules	25	"Mrs William has a healthy 30-year-old sister who has a healthy son and a daughter who..."	(Daughter, N/A)
Other	10	"Noah's mother died at age 72"	(Mother, N/A)
Wrong side	2	"...maternal paternal cousin..."	(Cousin, paternal)

[a]N/A: not applicable.

Partner's family annotations constitute the largest group of errors (approximately 23%), despite the use of the BERT-based paragraph filter. Without the filter, the number increased by more than 100%.

A large proportion (approximately 38%) of the errors fall into nonmention (approximately 21%) and nonerror (approximately 17%) categories. The distinction between these 2 classes is not always easy; for example, we classify an annotation of father in "[Patient's] father works in landscapin." as a nonmention, although it could well be interpreted as a nonerror. We believe this explains some of the differences in precision between the test set and the training set. On the training set, these 2 classes of FPs account for approximately 30% of total errors.

Lack of coreference contributes to approximately 18% of the errors. A closer look into the problem shows that many of these errors would require long-distance contexts (more than one sentence) to correctly resolve the references. A fairly common pattern is: "Patient's father suffers from.... His brother.... A sister...." The reference (brother/uncle) from the middle sentence can be resolved correctly fairly easily using a coreference resolution module. The reference form of the last sentence, however, is not explicit, and it requires context from both previous sentences. This points to an inherent limitation of our approach of applying the coreference resolution in the scope of sentence pairs.

Errors related to rule deficiencies account for approximately 14% of all FPs. The majority of these errors are related to the fact that our approach does not deal with enumerations, as only adjacent candidates are considered in rule-based processing (as shown in the example in Table 7). This problem can potentially be solved by incorporating sentence syntactic parsing. However, sentence parsing could introduce another algorithmic source of errors because errors in parsing caused by, for example, punctuation errors that are common in medical notes propagate into the downstream task.

Other errors refer to cases in which we were unable to determine whether the annotation is a nonerror or not. These notes refer to many different people by their first names, without explicitly stating who is the main subject of the note. A context external to the note might be necessary to produce the correct annotations.

Finally, sentence-level analysis shows that the family side heuristic works exceptionally well, producing very few errors

on rare occasions, such as with double cousins, as shown in Table 7.

A Closer Look at Coreference Resolution

To provide a better insight into the difficulty of incorporating a coreference resolution into a rule-based system, we compare the sentence-level analysis presented earlier with a similar experiment performed with the coreference resolution module.

We observe that the total number of errors is only 13 at the sentence level, but different errors are made. It is the larger variety of errors that contributes to lower precision. To provide an example of a common pattern, we can consider the following passage: "Her mother is healthy at age 63. Her father died at age 48 of COPD (Chronic Obstructive Pulmonary Disorder)." A system without the coreference resolution will produce correct *Mother* and *Father* annotations. A system with coreference resolution produces *Mother* and *Grandfather, Maternal* annotations, the latter being incorrect. Although it is incorrect, it is plausible both context-wise and grammatically. It is the accumulation of this type of mistake that negatively affects the precision score of the system with coreference resolution. We believe that a key takeaway is that in ambiguous cases (without explicit specification), choosing the patient as a reference point for a family relation is statistically safer than the coreference approach.

Missed FamilyMembers: FNs

Analyzing FNs within the test set is an inherently labor-intensive task, as it requires inspecting the entire FH note (to find the sentence-level evidence and identify why the system got it wrong). Our selective analysis indicates that a large proportion of the annotations missed by our system are related to the nonerrors detected in the exploration of FPs. For example, for the passage,

"This maternal aunt has three healthy children, but also had a daughter that died within the first few days of life secondary to hydrocephaly," our system provides an annotation (Cousin, Maternal). The gold standard requires an annotation of (Cousin, NA), which we assume relates to this particular text. As we believe the output of our system is correct, we classify it as a nonerror. At the same time, by correctly interpreting the sentence-level evidence, the system misses the gold standard annotation and the same nonerror penalizes both recall and precision.

Learning Points From FamilyMember Annotation

Our experiments with the FamilyMember annotation point to several high-level conclusions, which may be relevant for future work in this domain. First, the careful optimization of the system (error analysis on training data for debugging, choosing a more reliable set of relationships, and introducing BERT-based filtering) improves the overall performance of the system by more than 6% points, which we consider to be a fairly encouraging result. We are convinced that these results can be pushed even further with minor tweaks; however, it would be difficult to point to a specific thing that would drastically improve effectiveness if fixed. In addition, crafting additional rules that are very specific to the relatively few observed errors carry a risk of overfitting. In this sense, our approach has been taken relatively far for effective tuning.

Second, our experiments with coreference resolution demonstrate the intrinsic difficulty of configuring a *language understanding* component as an add-on to a rule-based system. We imagine that it is possible to come up with a much broader rule set that could take advantage of the coreference resolution. Nonetheless, not all context understanding can be solved with coreference resolution, especially for grammatically ambiguous cases or when deciding whether a matching surface form is an actual mention or a nonmention. This is even less useful when the coreference resolution is trained without task-specific context understanding. We believe that this points to a general limitation in rule-based approaches to the problem.

The use of pretrained neural LMs is the most viable path toward incorporating language understanding in the FamilyMember annotation. In our experiments, we demonstrate a simple BERT-based paragraph filtering approach, which improves the effectiveness of the final system. Its incorporation is easier than that of coreference resolution because we identify an isolated task (the interaction with the rest of the system is simple), which has task-specific training data (the method sees the contexts from a task-specific perspective). Nevertheless, a fully optimized ML baseline (BERT-CRF) does not outperform the rule-based approach.

In an ideal scenario, with thousands of training records, the FamilyMember annotation problem could be approached identically to that of Observations. However, with limited training data available for the task, such a model achieves effectiveness slightly lower than that of a rule-based system, as per our baseline experiments.

The third important takeaway relates to another advantage of a rule-based system, beyond the possibility of tuning F1 with very little training data. The rule-based system can generate conceptually correct annotations, regardless of the quality or completeness of the training data. We believe this is the reason we see so many nonerror FPs—our rules are conceptually sound. Therefore, the system will generate those outputs, even if the training data often miss a particular type of relationship. This means that rule-based systems, or their combinations, play an important role in creating annotated data sets that are needed to train deep learning approaches.

Overall Effectiveness of the System

We provide an overview of our best system's performance for FamilyMember and Observation annotations combined, compared with other approaches on the same data set, as shown in Table 8. As the FH notes collection is relatively new, most systems we compare against are those that participated in the 2019 N2C2 shared task (we selected the top 5 runs). We are aware that this comparison is not entirely fair, as we continued refining our system after the release of the test data, but it does put our results in perspective. For our best run, we present a combination of DAPT (ClinicalBERT) with ITPT (NCBI-disease) for Observation annotation and the best-performing system for the FamilyMember annotation. We

train the neural component for Observations with 5 different seeds and report average results with SDs.

The combined results demonstrate that the proposed system performs on par with top systems from the 2019 N2C2 shared task, with the exception of the Harbin Institute of Technology (HIT) team's approach, which achieves superior precision. We believe this is because of HIT proposing a model that jointly addresses both FamilyMember and Observation mentions via ML. It seems that their approach aligns better with the perks of this specific task (eg, annotating only diseases pertinent to specific FamilyMembers). In addition, our experiments with the BERT-CRF baseline for FamilyMember annotation indicate that the gap cannot be easily closed by simply using a state-of-the-art NER model for FamilyMember annotation. This also indicates that the key source of the difference in effectiveness between our best system and that of HIT is the HIT's feature of joint modeling of FamilyMember and Observation mentions.

Table 8. Comparison with other systems for both types of mentions combined[a].

Run	Precision	Recall	F1 score
Our best run, mean (SD)	79.60 (2.2)	83.64 (1.2)	81.63 (0.8)
HIT[b]	91.54	83.72	87.45
EZDI	80.90	83.65	82.25
MUSC[c]	78.90	83.84	81.30
NTTU[d]	80.43	80.93	80.68
UF[e]	79.69	79.20	79.44
N2C2[f] official median	—[g]	—	76.59
A1[h] [28]	65.01	88.92	75.10
A2[h] [28]	85.07	62.11	71.80

[a]National Natural Language Processing Clinical Challenges median is calculated from all valid runs participating in the original evaluation within the shared task.

[b]HIT: Harbin Institute of Technology.

[c]MUSC: Medical University of South Carolina.

[d]NTTU: National Taitung university.

[e]UF: University of Florida.

[f]N2C2: National Natural Language Processing Clinical Challenges.

[g]Not available.

[h]These are variants of the system described in the cited study.

In this study, we investigate the impact of a set of techniques (DATP and ITPT for disease annotation; rules and paragraph filtering for annotation of FMs) to improve the performance of a very simple yet reasonably effective baseline system (78.10 and 75.86 F1 scores for Observations and FamilyMembers, respectively, place it close to N2C2 median performance). Our experiments suggest that the proposed improvements, although subtle, generate a considerable cumulative effect, resulting in a final system performing at a close to state-of-the-art level. We also present a detailed error analysis for errors for the relatively less explored problem of annotating FMs in clinical notes.

Conclusions

We investigate the problem of detecting diseases and FM mentions in FH reports. We propose an approach that leverages state-of-the-art NER for disease mention detection, coupled with a hybrid method for FM mention detection. The hybrid method implements a rule-based approach combined with a text classifier to filter out irrelevant paragraphs from the reports (eg, pertaining to the patient partner's family).

Our approach achieved effectiveness close to the top 3 systems participating in the 2019 N2C2 FH extraction challenge, with only the top system outperforming it convincingly in terms of precision.

We believe that immediate improvements could be achieved by refining the rules used in the FM mention detection module. Nonetheless, alternative strategies, revolving around the use of semisuperised and distantly supervised learning, are closer to our research interests. A more encompassing approach toward improving performance would be a system that jointly models diseases and FMs, thereby improving cases that relate directly to the interplay between both entity types (eg, not annotating diseases of nonblood-related FMs).

In our future work, we will concentrate on applying FH extraction to a broader set of medical notes. This broader approach will not only cater to new use cases but will also allow for harnessing the FH-related knowledge scattered across other sections of EHRs.

Acknowledgments

This study was funded by the Commonwealth Science and Industrial Research Organization (CSIRO) Future Science Platform in Precision Medicine, Medical Decision Support Project. XD and SS were supported by the CSIRO's Data61 scholarship.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Implementation details. The configuration of our BERT model follows the original BERT-based model. In particular, our model is based on a bidirectional transformer with 768 hidden dimensions, 12 hidden layers, and 12 self-attention heads. The total number of parameters was approximately 110 million. We implemented our model using PyTorch and trained it using 1 RTX 2080 Ti graphics processing unit. As the training set size is small, iterating all instances once (1 epoch) takes less than 15 seconds. We adapted the early stop method, wherein the training will stop once there is no improvement (measured on the development set) during the last 5 consecutive epochs. The trained model that was most effective on the development set (measured using the F1 score) was used to evaluate the test set. BERT: Bidirectional Encoder Representations from Transformers.
[DOCX File , 9 KB - medinform_v9i4e24020_app1.docx]

References

1. 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 May 17;6(1):1-10. [doi: 10.1038/srep26094]
2. O'Malley AS, Draper K, Gourevitch R, Cross DA, Scholle SH. Electronic health records and support for primary care teamwork. J Am Med Assoc 2015 Mar;22(2):426-434 [FREE Full text] [doi: 10.1093/jamia/ocu029] [Medline: 25627278]
3. Dai X, Karimi S, Hachey B, Paris C. An effective transition-based model for discontinuous NER. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Online; 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online. [doi: 10.18653/v1/2020.acl-main.520]
4. Shen F, Liu S, Fu S, Wang Y, Henry S, Uzuner O, et al. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 National NLP Clinical Challenges (n2c2)/Open Health Natural Language Processing (OHNLP) competition. JMIR Med Inform 2021 Jan 27;9(1):e24008 [FREE Full text] [doi: 10.2196/24008] [Medline: 33502329]
5. 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]
6. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. Journal of Biomedical Informatics 2018 Jan;77:34-49 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.011] [Medline: 29162496]
7. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the AMIA Symposium 2001. 2001 Presented at: AMIA Symposium; 2001; Washington, DC.
8. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010 May 01;17(3):229-236. [doi: 10.1136/jamia.2009.002733]
9. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. 2008 Presented at: Pacific Symposium on Biocomputing; 2008; Kohala Coast, Hawaii. [doi: 10.1142/9789812776136_0062]
10. 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 2018 Mar 07;17(01):128-144. [doi: 10.1055/s-0038-1638592]
11. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform 2014;47:1-10. [doi: 10.1016/j.jbi.2013.12.006]
12. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Presented at: Advances in neural information processing systems; 2013; Lake Tahoe, Nevada.
13. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized word representations. 2018 Presented at: 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2018; New Orleans, Louisiana p. 2227-2237. [doi: 10.18653/v1/n18-1202]
14. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. A survey of word embeddings for clinical text. J Biomed Inform 2019 Dec;4:100057. [doi: 10.1016/j.yjbinx.2019.100057]
15. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019 Presented at: Clinical Natural Language Processing Workshop; 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/w19-1909]
16. Huang K, Altosaar J, Ranganath R. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. Cornell University. 2019. URL: https://arxiv.org/abs/1904.05342 [accessed 2020-12-01]

17. Zhang Y, Li HJ, Wang J, Cohen T, Roberts K, Xu H. Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes. 2018 Presented at: AMIA Summits on Translational Science; 2018; San Francisco, California.
18. Wang B, Lu W. Combining spans into entities: A neural two-stage approach for recognizing discontiguous entities. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1644]
19. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al. BioCreative/OHNLP Challenge 2018. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology,Health Informatics. 2018 Presented at: 2018 ACM International Conference on Bioinformatics, Computational Biology,Health Informatics; 2018; Washington, DC. [doi: 10.1145/3233547.3233672]
20. Polubriaginof F, Tatonetti NP, Vawdrey DK. An assessment of family history information captured in an electronic health record. 2015 Presented at: AMIA Annual Symposium; 2015; San Francisco, California.
21. Lewis N, Gruhl D, Yang H. Extracting Family History Diagnosis from Clinical Texts. 2011 Presented at: ISCA 3rd International Conference on Bioinformatics and Computational Biology; 2011; New Orleans, Luisiana.
22. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB. Automated extraction of family history information from clinical notes. In: AMIA Annual Symposium Proceedings 2014. 2014 Presented at: AMIA Annual Symposium; 2014; Washington, DC.
23. 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(S10):1-6. [doi: 10.1186/s12911-019-0995-5]
24. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. 2006 Presented at: AMIA Annual Symposium; 2006; Washington, DC.
25. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. 2008 Presented at: AMIA Annual Symposium; 2008; Washington, DC.
26. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 2006 Jul 26;6(1):1-9. [doi: 10.1186/1472-6947-6-30]
27. Cunningham H. GATE, a general architecture for text engineering. Comput Hum 2002;36(2):223-254. [doi: 10.3115/993268.993365]
28. Almeida JR, Matos S. Rule-based extraction of family history information from clinical notes. : Proceedings of the 35th Annual ACM Symposium on Applied Computing; 2020 Presented at: ACM Symposium on Applied Computing; 2020; Brno, Czech Republic p. 670-675. [doi: 10.1145/3341105.3374000]
29. Manning CD, Surdeanu M, Bauer J, Finkel JR, 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: 52nd annual meeting of the association for computational linguistics; 2014; Baltimore, Maryland. [doi: 10.3115/v1/p14-5010]
30. Dai H. Family member information extraction via neural sequence labeling models with different tag schemes. BMC Med Inform Decis Mak 2019 Dec 27;19(S10):1-12. [doi: 10.1186/s12911-019-0996-4]
31. 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 Dec 1;8(12):e21750. [doi: 10.2196/21750]
32. Zhan K, Peng W, Xiong Y, Fu H, Chen Q, Wang X, et al. Family history extraction using deep biaffine attention. JMIR Med Inform 2020 (forthcoming). [doi: 10.2196/preprints.23587]
33. Chinchor N. The statistical significance of the MUC-4 results. In: Proceedings of the 4th conference on Message understanding. 1992 Presented at: Conference on Message Understanding; 1992; Stroudsburg, Pennsylvania. [doi: 10.3115/1072064.1072068]
34. Tjong Kim Sang EF, De Meulder F. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of the Seventh Conference on Natural Language Learning. 2003 Presented at: HLT-NAACL; 2003; Edmonton, Canada p. 2003. [doi: 10.3115/1119176.1119195]
35. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019 Presented at: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019; Hong Kong, China. [doi: 10.18653/v1/d19-1371]
36. Sutton C, McCallum A. An introduction to conditional random fields for relational learning. In: Introduction to Statistical Relational Learning. Cambridge, Massachusetts: MIT Press; 2007.
37. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2015 Presented at: 54th Annual Meeting of the Association for Computational Linguistics; 2015; Berlin, Germany. [doi: 10.18653/v1/p16-1162]

38. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: Advances in neural information processing systems; 2017; Long Beach, California. [doi: 10.5555/3295222.3295349]
39. Devlin J, Chang MW, 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. 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019; Minneapolis, Minnesota p. 4171-4186.
40. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016 Presented at: 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2016; San Diego, California. [doi: 10.18653/v1/n16-1030]
41. Dai X, Karimi S, Hachey B, Paris C. Using similarity measures to select pretraining data for NER. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019 Presented at: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019; Minneapolis, Minnesota. [doi: 10.18653/v1/n19-1149]
42. Gururangan S, Marasović A, Swayamdipta S, Lo K, Beltagy I, Downey D, et al. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online. [doi: 10.18653/v1/2020.acl-main.740]
43. Pruksachatkun Y, Phang J, Liu H, Htut PM, Zhang X, Pang RY, et al. Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online. [doi: 10.18653/v1/2020.acl-main.467]
44. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
45. Uzuner O, 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]
46. Pradhan S, Elhadad N, South BR, Martinez D, Christensen LM, Vogel A, et al. Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. 2013 Presented at: Conference and Labs of the Evaluation Forum; 2013; Valencia, Spain. [doi: 10.1007/978-3-642-40802-1_24]
47. Fast Coreference Resolution in SpaCy with Neural Networks. GitHub. URL: https://github.com/huggingface/neuralcoref [accessed 2020-12-01]
48. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and Clinical English Model Packages in the Stanza Python NLP Library. Cornell University. 2020. URL: https://arxiv.org/abs/2007.14640 [accessed 2020-12-01]
49. Lange L, Adel H, Strötgen J. Joint De-Identification and Concept Extraction in the Clinical Domain. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020 Presented at: 58th Annual Meeting of the Association for Computational Linguistics; 2020; Online. [doi: 10.18653/v1/2020.acl-main.621]
50. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform 2015 Oct;57:28-37. [doi: 10.1016/j.jbi.2015.07.010]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory
CRF: Conditional Random Field
CSIRO: Commonwealth Science and Industrial Research Organization
DAPT: domain-adaptive pretraining
EHR: electronic health record
FH: family history
FM: family member
FN: false negative
FP: false positive
HIT: Harbin Institute of Technology
i2b2: Integrating Biology and the Bedside
IE: information extraction
ITPT: intermediate-task pretraining
LM: language model
ML: machine learning
N2C2: National Natural Language Processing Clinical Challenges

NCBI: National Center for Biotechnology Information
NER: named entity recognition
NLP: natural language processing
REX: Regenstrief Data Extraction
ShARe-CLEF: Shared Annotated Resources-Conference and Labs of the Evaluation Forum
TP: true positive

Edited by Y Wang, F Shen; submitted 31.08.20; peer-reviewed by Y Huang, D Mahajan; comments to author 03.11.20; revised version received 23.12.20; accepted 02.03.21; published 30.04.21.

Please cite as:
Rybinski M, Dai X, Singh S, Karimi S, Nguyen A
Extracting Family History Information From Electronic Health Records: Natural Language Processing Analysis
JMIR Med Inform 2021;9(4):e24020
URL: https://medinform.jmir.org/2021/4/e24020
doi:10.2196/24020
PMID:33664015

Original Paper

Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation

Yuanlin Yang[1,2*], MBA, BBA; Dehua Li[2,3*], MPhil, MBBS

[1]Department of Logistics Management, West China Second University Hospital, Sichuan University, Chengdu, China

[2]Key Laboratory of Obstetric and Gynecologic and Pediatric Disease and Birth Defects of Ministry of Education, Sichuan University, Chengdu, China

[3]Quality Assessment Office, Nursing Department, West China Second University Hospital, Sichuan University, Chengdu, China

[*]all authors contributed equally

Corresponding Author:
Dehua Li, MPhil, MBBS
Quality Assessment Office, Nursing Department
West China Second University Hospital
Sichuan University
No 20, Section 3, Renmin South Road
Wuhou District
Chengdu
China
Phone: 86 1 808 684 1792
Email: 562372162@qq.com

Abstract

Background: Big data technology provides unlimited potential for efficient storage, processing, querying, and analysis of medical data. Technologies such as deep learning and machine learning simulate human thinking, assist physicians in diagnosis and treatment, provide personalized health care services, and promote the use of intelligent processes in health care applications.

Objective: The aim of this paper was to analyze health care data and develop an intelligent application to predict the number of hospital outpatient visits for mass health impact and analyze the characteristics of health care big data. Designing a corresponding data feature learning model will help patients receive more effective treatment and will enable rational use of medical resources.

Methods: A cascaded depth model was successfully implemented by constructing a cascaded depth learning framework and by studying and analyzing the specific feature transformation, feature selection, and classifier algorithm used in the framework. To develop a medical data feature learning model based on probabilistic and deep learning mining, we mined information from medical big data and developed an intelligent application that studies the differences in medical data for disease risk assessment and enables feature learning of the related multimodal data. Thus, we propose a cascaded data feature learning model.

Results: The depth model created in this paper is more suitable for forecasting daily outpatient volumes than weekly or monthly volumes. We believe that there are two reasons for this: on the one hand, the training data set in the daily outpatient volume forecast model is larger, so the training parameters of the model more closely fit the actual data relationship. On the other hand, the weekly and monthly outpatient volume is the cumulative daily outpatient volume; therefore, errors caused by the prediction will gradually accumulate, and the greater the interval, the lower the prediction accuracy.

Conclusions: Several data feature learning models are proposed to extract the relationships between outpatient volume data and obtain the precise predictive value of the outpatient volume, which is very helpful for the rational allocation of medical resources and the promotion of intelligent medical treatment.

(JMIR Med Inform 2021;9(4):e19055) doi:10.2196/19055

KEYWORDS

deep learning; data mining; medical big data; model building

Introduction

Over the past two decades, there has been dramatic growth in the amount of data being generated in many areas worldwide, including health care data, sensor data, various types of user-generated data, internet data, and financial company data. Big data is emerging as the amount of data in every field grows; however, "big data" is an abstract concept that does not simply mean a large collection of data. Big data has some features that are different from data sets, and its characteristics differ from those of massive data and large data sets. Research studies examining concepts such as the Internet of Things and wearable technology have helped reduce the cost of real-time monitoring of human health, which has driven development in this industry [1]. Big data technology provides unlimited potential for efficient storage, processing, querying, and analysis of medical data. Technologies such as deep learning and machine learning simulate human thinking, assist physicians in diagnosis and treatment, provide personalized health services, and promote intelligent processes of health care applications. The development and application of the Internet of Things, wireless networks, the internet, cloud computing technology, etc., provide guarantees for the analysis, processing, and transmission of big data. In short, in the field of health care, the rapid development of big data analysis, wearable technology, artificial intelligence, Kyrgyz computing, supercomputing technology, etc., all provide possibilities for the realization and development of smart medical applications [2,3].

Medical health data is multimodal, complex data that continues to grow rapidly and contains a wealth of information. The challenges associated with medical health data include how to quickly and accurately collect medical health data and how to efficiently use high-speed networks to reliably and efficiently transmit medical health data [4]. Other challenges include the use artificial intelligence–related machine learning and deep learning techniques to extract useful information from health medical big data and the development of intelligent applications for medical staff and ordinary people. In this paper, our aims included analyzing health care data, addressing intelligent application–related issues, predicting the number of hospital outpatient visits for mass health impacts, and analyzing the characteristics of health care big data. Designing a corresponding data feature learning model will help patients receive more effective treatment and enable rational use of medical resources.

Methods

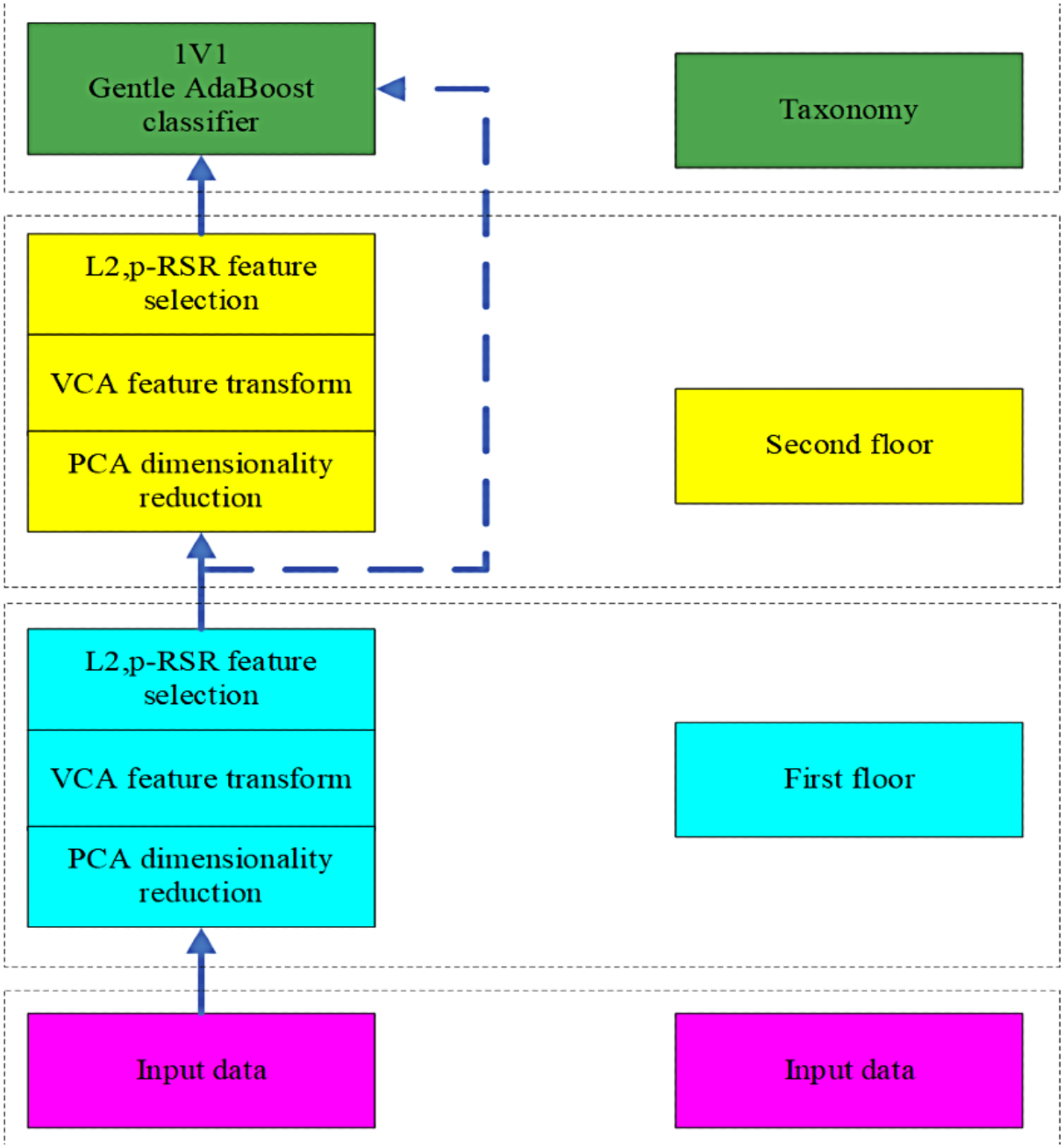
Cascaded Deep Learning Model

Model Framework

Deep learning is a process of feature extraction and combination. Through multilayer nonlinear operation combination, the model can abstract high-order semantic information as data. In practice, a cascaded multilayer operation is performed on the preprocessed raw data. Each layer consists of three sublayers: feature extraction, nonlinear transformation, and feature selection. The input of the initial layer is the pretreated original data, the input of the second layer is the output of the upper layer, and the output of the third layer is the final abstract representation of the data. In each layer, the representation characteristics of the data are extracted by feature transformation, which is generally a process of dimension increasing [5]. Compared with the input, the transformed features have certain representative characteristics; however, the number of features is much greater. Through feature selection, we reduce the dimension once, and at the same time, we choose the discriminant feature or the representation feature. Some feature selection methods are used to improve the regional adaptability of the model, such as max-pooling and mean-pooling operations in convolutional neural networks. Nonlinear transformation is usually performed in the middle of feature transformation and feature selection, and it is an important part of the framework. Nonlinear transformation can imitate the activation and inactivation of neurons. Another important application of nonlinear transformation is when linear transformation is used in the feature extraction. Multilayer linear transformation is still a linear transformation. Multilayer operation plays the same role as learning a linear transformation directly and does not play a role of layer-by-layer abstraction.

In this paper, to eliminate possible differences in measurement scales between different features, the original data were normalized by the maximum and minimum method. In the feature extraction stage, to obtain the structural information of the data, a method of feature transformation with little relationship to the domain was selected, that is, the vanishing component analysis (VCA) method was used to extract the polynomial characteristics of the data. Because the VCA method itself is a nonlinear transformation, the nonlinear operation cannot be carried out in the framework [6]. Because the dimension of the VCA input data cannot be too high, we used principal component analysis (PCA) to reduce the dimensions of the VCA input data (PCA reduces the dimension of the loss of information and can reduce the feature dimensions; therefore, it can also be regarded as a special feature selection method). In the final classification, we used a boosting algorithm, which is a method that can classify and select features; in the classification of the use of the features, we could use all the features learned at each level or only use the last level of features. At this point, our depth feature learning framework was formed, as shown in Figure 1.

Figure 1. The cascaded deep learning framework. IVI: innovation value intuitive; AdaBoost: adaptive boost; PCA: principal component analysis; RSR: regularized self-representation; VCA: vanishing component analysis.



Characteristic Learning

According to the existing cascaded depth learning framework, we implemented a specific depth model. Therefore, we needed to study and explain the training process of the model. The cascade depth model proposed in this paper is a multilayer structure. Between different layers, the output of the upper layer is the input of the next layer. Within the same layer, a PCA dimensionality reduction stage, VCA feature transformation stage, and $L_{2,p} - RSR$ feature selection stage, where RSR is regularized self-representation, are included. Each layer of the model can learn the output features of the current layer [7]. The abstract information of the different layers is different. The features of the lowest layer are closest to the original feature space. The high-level features can provide complementary information for the low-level features. We made full use of the characteristics learned from all levels and proposed an effective feature combination method. Finally, we used the boosting classifier based on the binary classification problem and extended its success to the multiclassification problem.

PCA Dimension Reduction Stage

VCA feature transformation requires stringent data space dimension control; therefore, it was necessary to reduce the dimension before the VCA feature transformation. There are two ways to reduce the dimension of PCA; one is to specify the reserved dimension directly, and the other is to set the ratio of an eigenvalue to the total sum of the eigenvalues. Proportion setting can theoretically control the retention percentage of data information; however, it is challenging to control the retained feature dimension using this method, especially when the feature dimension is difficult to control. In this model, the VCA

transform will produce a large number of features, and its input space needs to strictly control the feature dimension. Therefore, we directly set the reservation dimension of the PCA transform [8]. Because different dimensions retain different information, the performance of the model and the experimental results are affected, and this effect is not positively correlated with the retention dimension. Therefore, in our experiments, it was necessary to debug the PCA dimension.

VCA Feature Transformation Stage

The VCA method can map raw data into zero polynomial space, thus playing the role of feature extraction. Using this method, it is not necessary to know the domain of data usage or other prior knowledge because as long as the input space is a real matrix, we can learn its zero space polynomial transformation representation. This transformation method can not only extract the linear features in the data samples but can also extract the nonlinear features, that is, if the first-order polynomial contains the zero space of the data, the polynomial contains the linear information of the data. Other polynomials with different numbers can extract 2 or even higher order nonlinear characteristics of data. In this model, the use of VCA for feature transformation involves two specific problems: (1) polynomial number setting and initial feature dimension setting; (2) algorithm solution using singular value decomposition with the minimum setting.

$L_{2,p}$ – RSR Feature Selection Phase

VCA feature transformation will produce a large number of features, and we need to select features for many reasons. Moreover, feature selection is one of the reasons why the depth learning model is effective, as it can effectively select task-related features. The $L_{2,p}$ – *RSR* method proposed in this paper can not only effectively select features that play important roles in the linear representation of features but can also exclude the roles of singular samples due to the use of $L_{2,1}$ norm constraint loss terms. This method is based on the self-representation property of the feature space. Any matrix space data possesses this property; therefore, it is domain-independent, which meets the requirements of the generalization ability of our model. In this model, we only needed to set the P value of $L_{2,p}$ norm and regularization parameter λ value in the method. The input of the $L_{2,p}$ – *RSR* feature selection operation is the output space after VCA transformation, and the output is the output feature of the current layer of the depth model.

Boosting Classification and Feature Selection

After features are learned, it is necessary to classify them. There are many general classification methods, of which the nearest neighbor classifier is the simplest. Support vector machine (SVM) classifiers are also widely used in research and applications, and kernel-based SVM classifiers can also solve nonlinear problems. For this model, we used a classification method with a feature selection function: the Gentle Adaptive Boosting (AdaBoost) classifier based on a pile function. The Gentle AdaBoost algorithm based on a pile function can not only classify but can also select features from the feature space, that is, it can select one feature from each feature space and classify it, and it can select features by controlling the number of weak classifiers [9]. This is in good agreement with the framework proposed in this paper. The Gentle AdaBoost algorithm can not only play a classification role but can also perform feature selection. It can also be used only as a classification method or for feature selection. Next, we analyzed and implemented the boosting classifier based on a pile function.

Brief Summary of the Discrete AdaBoost Algorithm Based on a Pile Function

For the discrete AdaBoost algorithm based on a pile function, the inputs are training sample X and tag Y, and the output is the classifier model  .

Step 1: Weight matrix initialization: $w_i = 1/m, i = 1, \ldots, m$

Step 2: Repeat: $t = 1, 2, \ldots, T$

Step 3: For $d = 1, \ldots, n$, *do*: $(err^d, \delta^d, a^d, b^d) =$ 

Step 4: $feaId$ = ar g min (err_d), $(feaId, \delta, a, b) = (feaId, \delta^{feaId}, a^{feaId}, b^{feaId})$

Step 5: $f_t(x) = ah(x^{feaId} > \delta) + b$

Step 6: Update:  , $i = 1, 2, \ldots, m$, standardization of w makes 

Step 7: Repeating end

The Gentle AdaBoost algorithm based on the pile function follows the boosting algorithm framework and uses a simple classifier model:

in which the weak classifier f_m is defined as

The h function is the indicator function. represents the d dimension characteristic of the i sample. δ is the threshold (the so-called pile). a and b are parameters of the linear regression function. When learning the weak classifier, each feature of the sample learns a pile function based on the least squares, and the error value of the least squares is obtained and recorded; then, the corresponding feature is selected when the error value is the minimum. Therefore, (d, δ, a, b) can be obtained through the weighted least squares method:

After obtaining the weak classifier F_m, the weights of W are updated:

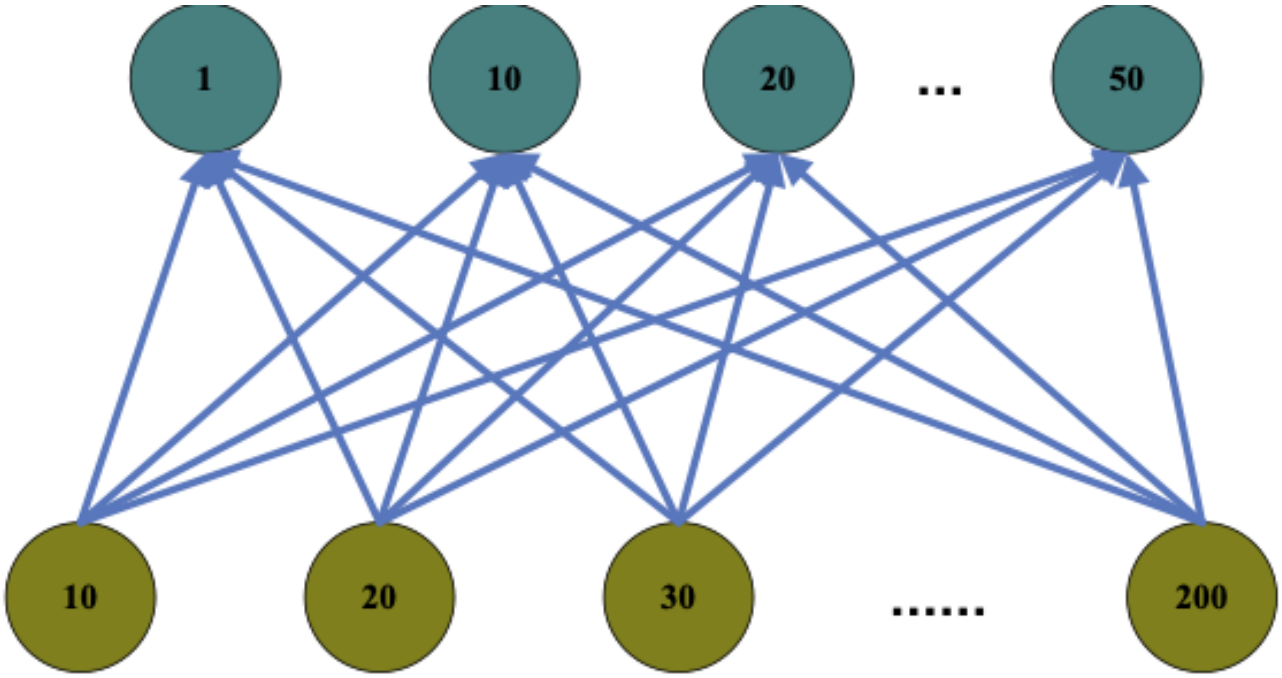
The F function is updated to $F = F + f$. The final classification result is *sign* $(F(x))$. The absolute value of the $F(x)$ value provides the credibility of the classification.

Feature Combination

Through the research and analysis of each stage of the cascade depth model, we successfully constructed a multilevel depth learning model and learned multilevel features. The different layer features provide different information. The underlying feature is the closest to the real information of the data and contains the most data information; however, it contains less semantic information and is not sufficiently abstract. When a model has two or three layers, it abstracts the features at different levels, including certain semantic information. When the number of layers is higher, the model contains a higher level of abstract information. We believe that each layer of features is useful; the underlying features ensure that the data information distortion is not too great, and the high-level features can provide the underlying features with complementary structure information that the underlying features do not possess [10]. If we can effectively combine each layer of features, we will obtain a good machine learning model. Using the combination of Gentle AdaBoost feature selection and classification functions, in this section, we propose an effective method of combining each layer of features and classifying them using the combination features [11-14]. The combination of the two levels of features is shown in Figure 2.

The classifier we used in the cascaded depth model is the Gentle AdaBoost classifier with a feature selection function. Therefore, we input the features of the different layers into the classifier algorithm to classify them and achieve the purpose of combining the features of each layer. The Gentle AdaBoost classifier is in the form of  . Assuming that the feature space of the first layer is X_1 and the feature space of the second layer is X_2, a strong classifier  is learned from the features of the first layer. Because the classification algorithm gives the current weight W of each sample after each weak classification, it can initialize the sample weight before F_2 training by using the weight distribution W of the sample at the end of F_1 after F_1 has been generated. Then, we learn the classifier  , and the final classification result is $F_1 + F_2$. F_1 uses only the first layer feature and F_2 uses only the second layer feature; however, the sample weight W is used in the F_1 process.

Figure 2. The combination of two levels of features used in our model.



Data Analysis

We used the Letter, Pendigit, and USPS data sets to conduct comparative experiments [15-18]. Specifically, the example, feature, and class labels are 20,000, 16, and 26 for the Letter data set, 10,992, 16, and 10 for the Pendigit data set, and 9298, 256, and 10 for the USPS data set, respectively. From Table 1, we can see that the data pair classification accuracy was low; we selected the lowest number, and the corresponding sequence was (31 41 45 125 128 157 164 173 304). The bandwidth of the first layer PCA was (4810 1316), that of the second layer was (358 11 13 15 20) and that of the third layer as (358 11 13 15). The classification accuracy of the selected data under one level of PCA is shown in Table 1. As can be seen from the table, the optimal value was PCA1=13. With a fixed PCA1 of 13, the second layer under different PCA2 values of classification accuracy is shown in Table 2; the optimal value was PCA2=11. Moreover. the third layer results are shown in Table 2, and the corresponding PCA3 was 13. In actual bandwidth settings, PCA reserve values can be set for several more groups because Letter data have a smaller sample size per class. Finally, each layer used the PCA reservation dimension (13 11 13) setting to obtain the total classification accuracy of Letter data under each layer. Tables 1-2 also show that the classification accuracy increases with the increase of the number of layers when the model classifies Letter data with attribute values. Figure 3 and Figure 4 show the classification accuracies of the Pendigit and USPS data sets at different layers, respectively.

Table 1. Classification accuracy of the first layer of Letter data with different PCA1 retention values (%).

Data serial number	PCA1[a] retention value				
	4	8	10	13	16
31	92.10	98.63	99.62	99.62	99.13
41	85.93	94.58	94.58	97.59	98.51
45	97.13	98.46	98.89	98.89	97.90
125	83.87	94.20	96.78	98.52	98.52
128	89.92	95.96	96.98	99.65	99.65
157	84.59	96.33	94.92	97.87	95.89
164	94.10	95.12	97.42	98.80	99.27
173	91.83	95.11	96.89	98.58	97.30
304	85.66	96.92	99.29	100.00	100.00
Average value	89.60	96.18	97.28	98.89	98.72

[a]PCA1: principal component analysis 1.

Table 2. Classification accuracy of Letter data under different PCA2 retention values (%).

Data serial number	PCA2[a] retention value					
	3	5	8	11	13	15
31	99.62	99.62	99.62	99.13	99.62	99.62
41	97.59	97.99	97.59	98.59	98.59	97.89
45	99.78	99.78	99.23	99.78	99.23	99.78
125	99.39	99.39	99.39	99.39	99.82	99.25
128	99.65	100.00	99.65	99.65	99.65	99.65
157	98.94	98.94	98.94	98.94	98.49	98.94
164	99.19	99.19	99.73	98.73	99.19	99.19
173	98.91	98.91	98.91	98.91	98.91	98.58
304	100.00	100.00	100.00	100.00	100.00	100.00
Average value	99.59	99.69	99.58	99.59	99.75	99.69

[a]PCA2 : principal component analysis 2.

Figure 3. Classification accuracy of the Pendigit data set at different layers.

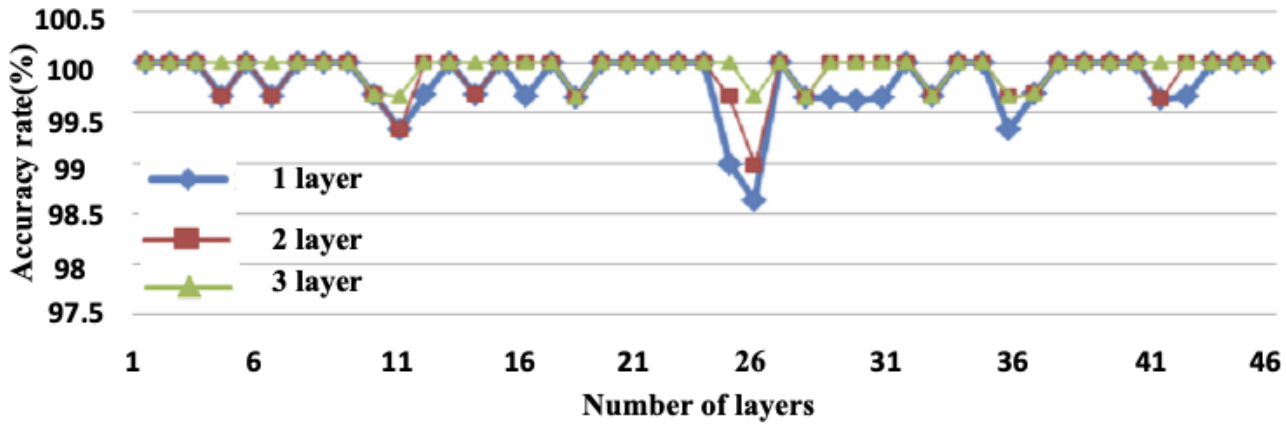
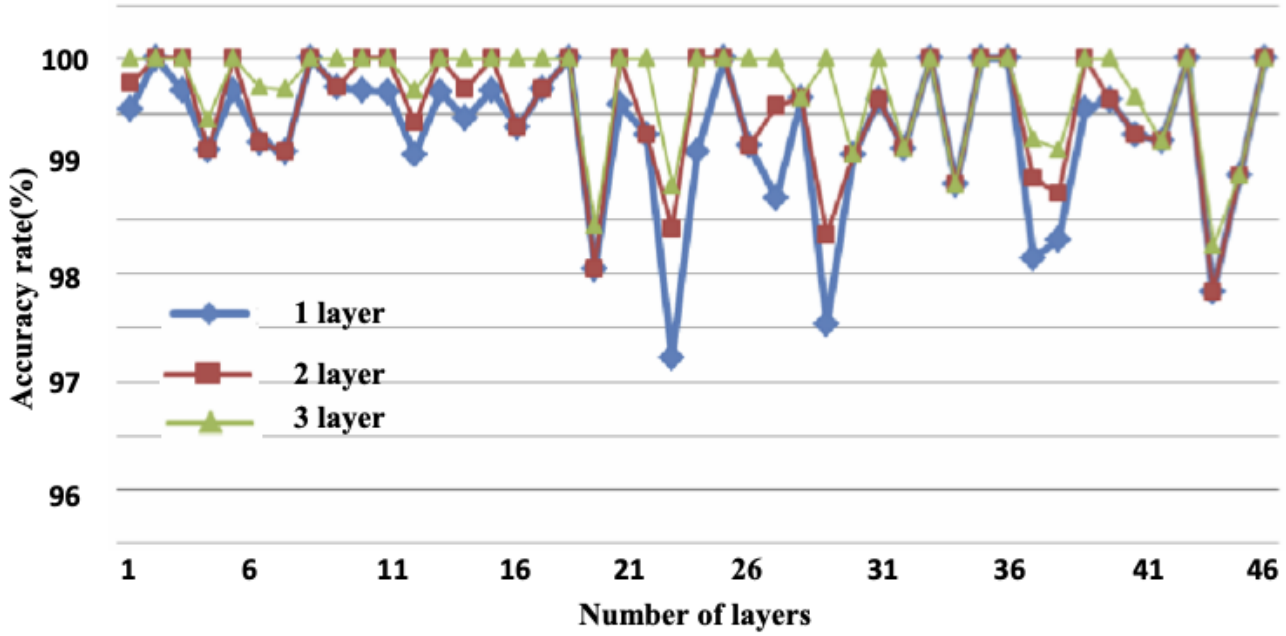


Figure 4. Classification accuracy of the USPS data set at different levels.



Prediction of Hospital Outpatient Volume Based on the RNN-RBM Network Model

Prediction Model Construction

Recurrent neural networks (RNNs) and restricted Boltzmann machine (RBM) networks have strong ability to predict time series. To make use of the advantages of these two network models to predict hospital outpatient volumes, the two networks were combined to form a depth structure network RNN-RBM model. This deep neural network can describe the time dependence of high-dimensional sequences.

Among the various deformations of RBM, there is a deformation called conditional RBM. Conditional RBM is different from standard RBM in that it adds two types of connections: one is an autoregressive connection, and the other is a hidden layer connection between the previous time step and the current time step. A conditional RBM can also be trained by a contrastive divergence (CD) algorithm. This structure can be used to handle time series data effectively; therefore, it can be used to solve time series problems. The RNN-RBM model in this paper can also refer to conditional RBM. and represent the offset vectors of the visible layer and the hidden layer in the RBM model at time t, and they are updated by the hidden unit u^{t-1} in the RNN model at time t – 1. Weight matrices W_{uv} and W_{uh} are used to connect the RNN network model and RBM network model. The bias vector can be expressed as follows:

where b_v and b_h are the initial offset values of the visible and hidden layers in the RBM network model. The RNN network model expands gradually with the time step and is used to generate the state of the hidden units in the RBM network model, which are based on the input layer $v^{(t)}$ and the hidden layer $u^{(t)}$ in the RNN network model. In this way, the hidden layer can only blame the activation function of the hidden unit.

$$u^{(t)} = sigmoid\ (b_u + W_{uu}u^{(t-1)} + W_{vu}v^{(t)})$$

From this, we can see that the overall process of the algorithm is:

1. The hidden unit in the RNN model is activated.
2. $u^{(t-1)}$ is used in Step 1 to update the bias values in the RBM network model.
3. The parameters are updated in the RBM network model.
4. The RBM output is used as the input of the prediction layer, and the parameters are initialized randomly.
5. The backpropagation (BP) method is used to fine-tune the model from top to bottom. Error values are propagated back to the RBM and RNN network models. The weight matrices w_{uv} and w_{uh} are updated, and the RNN network models are trained to predict.

Results

The experimental simulation data were obtained from real-world hospital outpatient volume data. In the data pretreatment stage, in this paper, we extracted the outpatient volume information of each department of the hospital and performed certain statistical processing methods. In the actual processing data, the outpatient data for hospital holidays were significantly reduced; to avoid the impact of these data, we deleted the related statistical processing methods.

To better satisfy the prediction function, these data were counted according to the three time intervals of day, week, and month. To better conform to the prediction model based on the depth neural network proposed in this paper, these data were processed and expressed as a data matrix. The matrix is as follows:

in which represents the outpatient volume data of department *n* in interval *T*.

In this paper, RNN and RBM neural networks were combined to form a deep-seated neural network model, and the model was used to predict the outpatient volume of the hospital. In the actual simulation, we selected the outpatient volume data of 15 important outpatient clinics as the input of the depth model, that is, the input layer of the model was set to 15. The number of hidden layer neurons in the RNN was set to 20, the number of hidden layer neurons in the RBM was set to 30, and the output of the predicted layer was 15.

Figure 5 and Figure 6 show the prediction results of the RNN-RBM model. In the simulation experiment, the outpatient volume data were trained in different intervals (daily outpatient volume, weekly outpatient volume, and monthly outpatient volume). From the simulation results, it can be seen that the forecast of the daily outpatient volume is closer to the real value than that of weekly outpatient volume and monthly outpatient volume. The depth model created in this paper is more suitable for the daily outpatient volume forecast, and we analyzed the causes of this phenomenon. We believe that there are two reasons for this result. On the one hand, the amount of training data in the daily outpatient volume forecast model is larger; therefore, the training parameters of the model more closely fit the actual data relationship. On the other hand, the weekly and monthly outpatient volume is the cumulative daily outpatient volume; therefore, the errors caused by the prediction will gradually accumulate, and the greater the interval, the lower the prediction accuracy.

In practical applications, medical managers often require more short-term outpatient volume forecasts. Because forecasting a shorter interval of outpatient volume can provide support for medical management, this method still has certain advantages in practical applications.

We compared the outpatient volume forecasting method based on the RNN-RBM depth model with existing popular forecasting algorithms, and the results are shown in Table 3. Here, the prediction algorithm based on the auto regressive integrated moving average (ARIMA) model, the BP neural network prediction algorithm, the radial basis function (RBF) neural network prediction algorithm, and the SVM algorithm were selected.

Compared with the current popular outpatient volume prediction algorithm, it can be readily concluded that the prediction algorithm based on the RNN-RBM depth model is superior to other current prediction algorithms for the daily outpatient volume, weekly outpatient volume, and monthly outpatient volume, and the prediction accuracy is relatively high.

Figure 5. Daily outpatient volume forecast results.

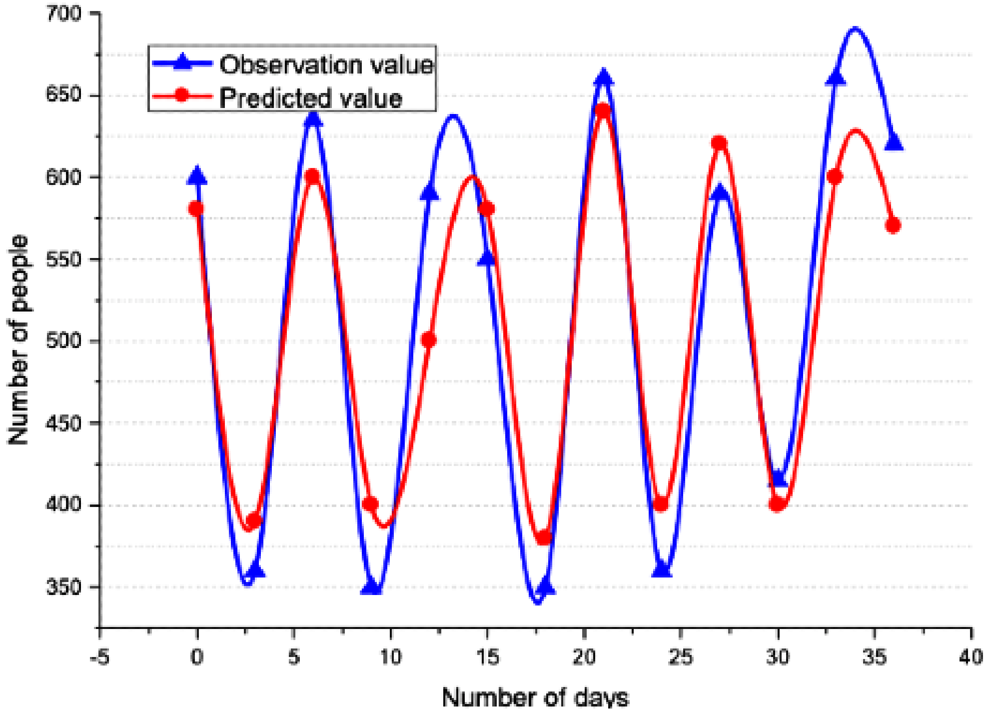


Figure 6. Weekly outpatient volume forecast results.

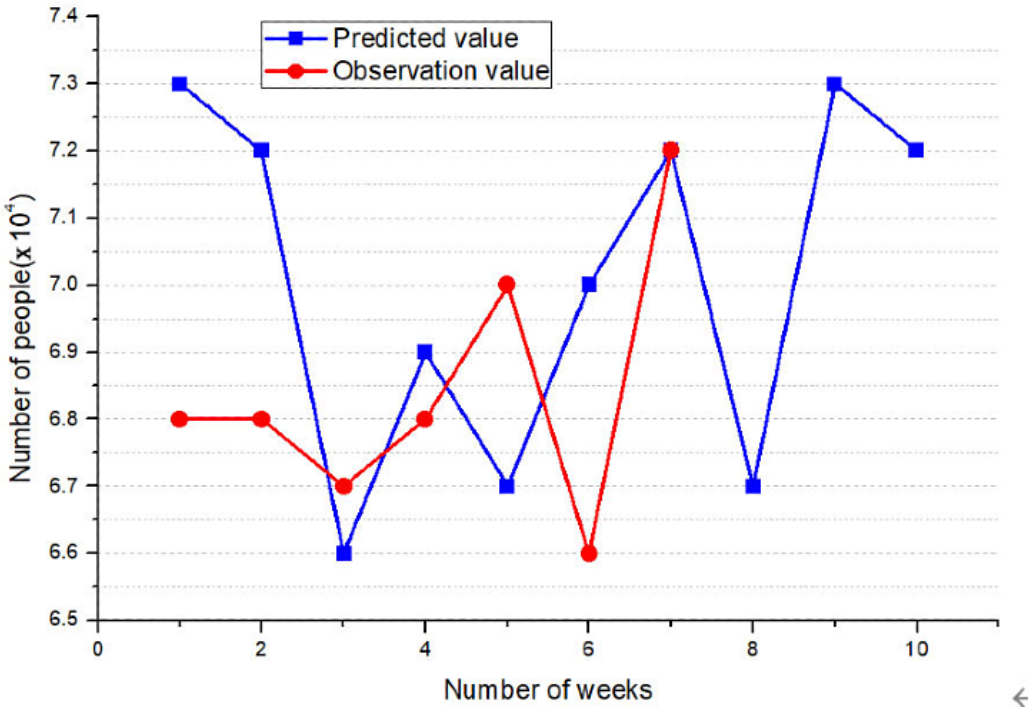


Table 3. Comparison of outpatient volume prediction methods.

Method and error	Daily outpatient volume	Weekly outpatient volume	Monthly outpatient volume
RNN-RBM[a]			
MAE[b]	24.83	2395.1	11966.6
MRE[c] (%)	3.9	4.6	5.9
ARIMA[d]			
MAE	37.9	3562.7	19525.5
MRE (%)	5.4	6.1	6.8
BP[e]			
MAE	53.48	4145.5	2549.5
MRE (%)	7.3	8.9	9.6
SVM[f]			
MAE	44.3	3796.9	2647.8
MRE (%)	6.4	7.9	9.4
RBF[g]			
MAE	38.9	3586.3	22978
MRE (%)	5.7	7.1	9

[a]RNN-RBM: recurrent neural network-restricted Boltzmann machine.

[b]MAE: mean absolute error.

[c]MRE: mean relative error.

[d]ARIMA: auto regressive integrated moving average.

[e]BP: backpropagation.

[f]SVM: support vector machine.

[g]RBF: radial basis function.

Discussion

Big data in the field of health care is an integral part of the strategic layout of national big data, and the analysis and mining of valuable information is also related to the development of national health care. At present, the problems that must be solved in the analysis and application of health care data include timely and accurate collection and acquisition of health care data as well as efficient use of high-speed networks for reliable transmission of health care–related digital, image, voice, and other information. Machine learning and in-depth learning technology related to artificial intelligence can be used to mine useful information from health care–related big data and develop intelligent applications for medical staff and ordinary people. In this paper, we studied a feature learning model of medical health data based on probabilistic and in-depth learning mining, mining information from medical big data and addressing intelligent application–related problems, and we studied the differences between medical risk assessment–related data and general big data, multimodal data feature representation, and learning–related content. Several data feature learning models are proposed to extract the relationship between the outpatient volume data and to obtain precise predictive value of outpatient volume, which is very helpful to the rational allocation of medical resources and the promotion of intelligent medical treatment.

Acknowledgments

The authors acknowledge the Soft Science Project of the Sichuan Science and Technology Department (Grant: 2017ZR0169).

Conflicts of Interest

None declared.

References

1. Nie L, Zhao Y, Akbari M, Shen J, Chua T. Bridging the Vocabulary Gap between Health Seekers and Healthcare Knowledge. IEEE Trans Knowl Data Eng 2015 Feb 1;27(2):396-409. [doi: 10.1109/tkde.2014.2330813]
2. Lewis D, Pluye P, Rodriguez C, Grad R. Mining reflective continuing medical education data for family physician learning needs. J Innov Health Inform 2016 Apr 06;23(1):834-440 [FREE Full text] [doi: 10.14236/jhi.v23i1.834] [Medline: 27348489]
3. Shen D, Zhang D, Young A, Parvin B. Machine Learning and Data Mining in Medical Imaging. IEEE J Biomed Health Inform 2015 Sep;19(5):1587-1588. [doi: 10.1109/jbhi.2015.2444011] [Medline: 26574616]
4. Wong KKL, Wang D, Fong S, Ng EYK. A Special Section on Advanced Computing Techniques for Machine Learning and Data Mining in Medical Informatics. J Med Imaging Hlth Inform 2016 Aug 01;6(4):1052-1055. [doi: 10.1166/jmihi.2016.1800]
5. Van Eaton EG, Devlin AB, Devine EB, Flum DR, Tarczy-Hornoch P. Achieving and sustaining automated health data linkages for learning systems: barriers and solutions. EGEMS (Wash DC) 2014;2(2):1069 [FREE Full text] [doi: 10.13063/2327-9214.1069] [Medline: 25848606]
6. Sun W, Cai Z, Li Y, Liu F, Fang S, Wang G. Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. JHE 2018 Apr 08;2018:1-9. [doi: 10.1155/2018/4302425]
7. Kusy M, Zajdel R. Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification. Appl Intell 2014 Aug 6;41(3):837-854. [doi: 10.1007/s10489-014-0562-9]
8. Bashir MEA, Shon HS, Lee DG, Kim H, Ryu KH. Real-Time Automated Cardiac Health Monitoring by Combination of Active Learning and Adaptive Feature Selection. KSII TIIS 2013 Jan 29;7(1):99-118. [doi: 10.3837/tiis.2013.01.007]
9. Henriques R, Antunes C, Madeira SC. Generative modeling of repositories of health records for predictive tasks. Data Min Knowl Disc 2014 Nov 12;29(4):999-1032. [doi: 10.1007/s10618-014-0385-7]
10. Bos JW, Lauter K, Naehrig M. Private predictive analysis on encrypted medical data. J Biomed Inform 2014 Aug 14;50:234-243 [FREE Full text] [doi: 10.1016/j.jbi.2014.04.003] [Medline: 24835616]
11. Wu S. A Traffic Motion Object Extraction Algorithm. Int. J. Bifurcation Chaos 2016 Jan 14;25(14):1540039. [doi: 10.1142/S0218127415400398]
12. Wu S, Wang M, Zou Y. Research on internet information mining based on agent algorithm. Future Generation Computer Systems 2018 Sep;86:598-602. [doi: 10.1016/j.future.2018.04.040]
13. Ke Q, Wu S, Wang M, Zou Y. Evaluation of Developer Efficiency Based on Improved DEA Model. Wireless Pers Commun 2018 Feb 7;102(4):3843-3849. [doi: 10.1007/s11277-018-5415-0]
14. Wu S, Wang M, Zou Y. Sewage information monitoring system based on wireless sensor. Desalination Water Treat 2018 Jul;121:73-83. [doi: 10.5004/dwt.2018.22362]
15. Cour T, Sapp B, Taskar B. Learning from Partial Labels. The Journal of Machine Learning Research 2011 Jul;12:1501-1536 [FREE Full text] [doi: 10.5555/1953048.2021049]
16. Chen Y, Patel V, Chellappa R, Phillips P. Ambiguously Labeled Learning Using Dictionaries. IEEE Trans.Inform.Forensic Secur 2014 Dec;9(12):2076-2088. [doi: 10.1109/TIFS.2014.2359642]

17. Liu L, Dietterich T. A conditional multinomial mixture model for superset label learning. In: Bartlett P, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 25. Cambridge, MA: MIT Press; 2012:557-565.
18. Zhang ML, Zhou BB, Liu XY. Partial Label Learning via Feature-Aware Disambiguation. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug Presented at: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August, 2016; San Francisco, CA p. 1335-1344.

Abbreviations

AdaBoost: adaptive boosting
ARIMA: auto regressive integrated moving average
BP: backpropagation
CD: contrastive divergence
PCA: principal component analysis
RBF: radial basis function
RBM: restricted Boltzmann machine
RNN: recurrent neural network
RSR: regularized self-representation
SVM: support vector machine
VCA: vanishing component analysis

Edited by Z Du, K Kalemaki, H Li; submitted 02.04.20; peer-reviewed by E Marcs, H Wang, J Fu; comments to author 14.04.20; revised version received 08.05.20; accepted 08.05.20; published 08.04.21.

Please cite as:
Yang Y, Li D
Medical Data Feature Learning Based on Probability and Depth Learning Mining: Model Development and Validation
JMIR Med Inform 2021;9(4):e19055
URL: https://medinform.jmir.org/2021/4/e19055
doi:10.2196/19055
PMID:33830067

Original Paper

Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study

Jenna M Reps[1], BSc, MSc, PhD; Chungsoo Kim[2], PharmD; Ross D Williams[3], MSc; Aniek F Markus[3], BSc, MSc; Cynthia Yang[3], BSc, MSc; Talita Duarte-Salles[4], MPH, PhD; Thomas Falconer[5], BSc, MSc; Jitendra Jonnagaddala[6], MIS, PhD; Andrew Williams[7], PhD; Sergio Fernández-Bertolín[4], MSc; Scott L DuVall[8], PhD; Kristin Kostka[9], MPH; Gowtham Rao[1], MD, PhD; Azza Shoaibi[1], PhD; Anna Ostropolets[5], MD; Matthew E Spotnitz[5], MPH, MD; Lin Zhang[10,11], PhD; Paula Casajust[12], BSc, MSc; Ewout W Steyerberg[13,14], MSc, PhD; Fredrik Nyberg[15], MPH, MD, PhD; Benjamin Skov Kaas-Hansen[16,17], MSc, MD; Young Hwa Choi[18], MD, PhD; Daniel Morales[19], PhD, MBChB; Siaw-Teng Liaw[6], PhD; Maria Tereza Fernandes Abrahão[20], PhD; Carlos Areia[21], MSc, PT; Michael E Matheny[22], MD, MPH, MS; Kristine E Lynch[8], PhD; María Aragón[4], MSc; Rae Woong Park[23], MD, PhD; George Hripcsak[5], MD, MS; Christian G Reich[9], MD, PhD; Marc A Suchard[24], MD, PhD; Seng Chan You[23], MD, MS; Patrick B Ryan[1], PhD; Daniel Prieto-Alhambra[25], MD, PhD; Peter R Rijnbeek[3], PhD

[1]Janssen Research & Development, Titusville, NJ, United States
[2]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea
[3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, Netherlands
[4]Fundacio Institut Universitari per a la recerca a l'Atencio Primaria de Salut Jordi Gol i Gurina, Barcelona, Spain
[5]Department of Biomedical Informatics, Columbia University, New York, NY, United States
[6]School of Public Health and Community Medicine, University of New South Wales, Sydney, Australia
[7]Tufts Institute for Clinical Research and Health Policy Studies, Boston, MA, United States
[8]Department of Veterans Affairs, University of Utah, Salt Lake City, UT, United States
[9]Real World Solutions, IQVIA, Cambridge, MA, United States
[10]Melbourne School of Public Health, The University of Melbourne, Victoria, Australia
[11]School of Public Health, Peking Union Medical College, Beijing, China
[12]Department of Real-World Evidence, Trial Form Support, Barcelona, Spain
[13]Department of Public Health, Erasmus University Medical Center, Rotterdam, Netherlands
[14]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, Netherlands
[15]School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[16]Clinical Pharmacology Unit, Zealand University Hospital, Roskilde, Denmark
[17]NNF Centre for Protein Research, University of Copenhagen, Copenhagen, Denmark
[18]Department of Infectious Diseases, Ajou University School of Medicine, Suwon, Republic of Korea
[19]Division of Population Health and Genomics, University of Dundee, Dundee, United Kingdom
[20]Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil
[21]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom
[22]Department of Veterans Affairs, Vanderbilt University, Nashville, TN, United States
[23]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea
[24]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, United States
[25]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Jenna M Reps, BSc, MSc, PhD
Janssen Research & Development
1125 Trenton Harbourton Rd
Titusville, NJ
United States
Phone: 1 732 715 6300

Email: jreps@its.jnj.com

Abstract

Background: SARS-CoV-2 is straining health care systems globally. The burden on hospitals during the pandemic could be reduced by implementing prediction models that can discriminate patients who require hospitalization from those who do not. The COVID-19 vulnerability (C-19) index, a model that predicts which patients will be admitted to hospital for treatment of pneumonia or pneumonia proxies, has been developed and proposed as a valuable tool for decision-making during the pandemic. However, the model is at high risk of bias according to the "prediction model risk of bias assessment" criteria, and it has not been externally validated.

Objective: The aim of this study was to externally validate the C-19 index across a range of health care settings to determine how well it broadly predicts hospitalization due to pneumonia in COVID-19 cases.

Methods: We followed the Observational Health Data Sciences and Informatics (OHDSI) framework for external validation to assess the reliability of the C-19 index. We evaluated the model on two different target populations, 41,381 patients who presented with SARS-CoV-2 at an outpatient or emergency department visit and 9,429,285 patients who presented with influenza or related symptoms during an outpatient or emergency department visit, to predict their risk of hospitalization with pneumonia during the following 0-30 days. In total, we validated the model across a network of 14 databases spanning the United States, Europe, Australia, and Asia.

Results: The internal validation performance of the C-19 index had a C statistic of 0.73, and the calibration was not reported by the authors. When we externally validated it by transporting it to SARS-CoV-2 data, the model obtained C statistics of 0.36, 0.53 (0.473-0.584) and 0.56 (0.488-0.636) on Spanish, US, and South Korean data sets, respectively. The calibration was poor, with the model underestimating risk. When validated on 12 data sets containing influenza patients across the OHDSI network, the C statistics ranged between 0.40 and 0.68.

Conclusions: Our results show that the discriminative performance of the C-19 index model is low for influenza cohorts and even worse among patients with COVID-19 in the United States, Spain, and South Korea. These results suggest that C-19 should not be used to aid decision-making during the COVID-19 pandemic. Our findings highlight the importance of performing external validation across a range of settings, especially when a prediction model is being extrapolated to a different population. In the field of prediction, extensive validation is required to create appropriate trust in a model.

(JMIR Med Inform 2021;9(4):e21547) doi:10.2196/21547

KEYWORDS

external validation; transportability; COVID-19; prognostic model; prediction; C-19; modeling; datasets; observation; hospitalization; bias; risk; decision-making

Introduction

Background

The novel coronavirus SARS-CoV-2, which causes COVID-19, is quickly spreading throughout the world and burdening health care systems worldwide [1]. Numerous prediction models are being developed and released to the public to aid decision-making during the pandemic [2]. Many of these models aim to inform people of their risk of developing severe outcomes due to COVID-19 [3-5]. A recent systematic review found that all the then-published models suffered from high risk of bias along with one or more limitations, including small data sets used to develop the models and lack of external validation [2].

The COVID-19 vulnerability (C-19) index [5] is an example of a prognostic model developed to identify people susceptible to severe outcomes during COVID-19 infection. The model is potentially valuable because it aims to predict hospitalization risk in the general population [2]. At the time of the study, a paper on the model was available as a preprint [5], and the model itself was publicly available at a website [6]. The C-19 index aims to predict which patients will require hospitalization due to pneumonia (or proxies for pneumonia) within 3 months. The model was developed using retrospectively collected Medicare data (patients aged 65 years or older) that did not include patients with COVID-19.

Objectives

In this paper, we aim to show the importance of external validation and demonstrate the feasibility, during times of urgency, of using a collaborate network for this purpose. We chose to demonstrate this with the C-19 index because it is available as a commercial product to the public, prior to being peer-reviewed, as a model that can predict COVID-19 severity, but it has not undergone any external validation. It is unknown whether this model is currently being used for medical decision-making, but it has been advertised as a decision-making tool. However, the process illustrated in this paper and the lessons learned are applicable to any COVID-19 prediction model. Furthermore, the C-19 index model was developed using non–COVID-19 data, and there is no guarantee that a model trained on Medicare patients who do not have COVID-19 will perform similarly or even adequately in patients with COVID-19. Research has shown that there is high risk of bias for a model that lacks external validation [7]. In addition, it is

recommended to assess the knowledge of a model's reproducibility and transportability before it is used clinically [8]. Models must be reliable, as poor predictions can be detrimental to decision-making [2].

The Observational Health Data Science and Informatics (OHDSI) collaboration is a group of researchers who are collaborating to develop best practices for analyzing observational health care data [9]. OHDSI has developed a framework that enables timely validation of prediction models across a large number of data sets worldwide [10]. The OHDSI network currently contains large COVID-19 cohorts from the United States, Europe, and Asia. In this study, we aim to demonstrate the importance of performing external validation of a model before its predictions can be trusted. As a case study, we chose to investigate the predictive performance of the C-19 index when applied to COVID-19 data from databases across the world. This study provides information about the suitability of using the C-19 index model to aid decision-making during the COVID-19 pandemic.

Methods

Existing C-19 Index Models

Three models were developed in the C-19 index paper [5]. The simplest model was a logistic regression with a limited number of predictors: age, sex, hospital usage, 11 comorbidities, and their age interactions. The other two models were less parsimonious gradient boosting machines with more than 500 variables. Only one of these gradient boosting machine models was reported. Withholding a model results in noncompliance with the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) statement [11] and makes external validation impossible. In this paper, we chose to evaluate the simple logistic regression model, recognizing that COVID-19 prediction models are urgently needed worldwide and that parsimonious models are more readily implemented across health care settings.

Data Source

Electronic medical records (EMRs) and administrative claims databases from primary care and secondary care systems containing patients from Australia, Japan, the Netherlands, Spain, South Korea, and the United States were analyzed in a distributed network, as detailed in Multimedia Appendix 1, Table S1. Of these data sets, 5 contained COVID-19 cases and 9 did not. All data sets used in this paper were mapped into the OHDSI Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [12]. The OMOP-CDM was developed to provide researchers with diverse data sets with a standard database structure. This enables analysis code and software to be shared among researchers, which facilitates external validation of prediction models. Deidentified or pseudonymized data were obtained from routinely collected records from clinical practice. Analyses were performed using the following databases: the Australia Electronic Practice–Based Research Network (AU-ePBRN) (linked primary and secondary care database from Australia); Japanese Medical Data Center (JMDC) (Japanese claims); Integrated Primary Care Information (IPCI) (primary care EMR from the Netherlands); Information System for Research in Primary Care (SIDIAP) (primary care EMR from Spain); Ajou University School of Medicine (AUSOM) and Health Insurance Review and Assessment (HIRA) (EMR and claims database, respectively, from South Korea); Commercial Claims and Encounters (CCAE), ClinFormatics, Medicare (MDCR), Medicaid (MDCD) (US claims databases), Optum EHR, Department of Veterans Affairs (VA), Columbia University Irving Medical Center (CUIMC) and Tufts Medical Center Research Data Warehouse (TRDW) (US EMRs). All analyses were conducted locally in a distributed network in which the analysis code was sent to participating sites and only aggregate summary statistics were returned, with no sharing of patient-level data between organizations.

Consent to Publish

Each site obtained institutional review board approval for the study or used deidentified data; therefore, the study was determined not to be human subject research. Informed consent was not necessary at any site.

Participants

The purpose of the C-19 index is to identify which patients with COVID-19 are more likely to require hospitalization due to severe complications. The C-19 index model was developed using non–COVID-19 data; therefore, we externally validated it in (1) COVID-19 cohorts, to see how well the model transports to patients it is being advertised for, and (2) non–COVID-19 cohorts, to see how well the model transports to patients similar to those used to develop it.

We chose to investigate the performance of the model when applied to patients with an outpatient or emergency department (ED) visit with initial symptoms. We chose this approach because it mimics the situation in which patients first seek treatment or medical advice due to developing symptoms or testing positive for COVID-19 (or influenza).

For the external validation using COVID-19 data, patients were included in the target population if they satisfied the criteria below:

- Presenting at an outpatient or ED visit with COVID-19 (COVID-19 was identified by a diagnosis code for SARS-COV-2 or a positive test for SARS-COV-2 that was recorded after January 1, 2020)
- Aged ≥18 years during the outpatient or ED visit
- ≥365 days of observation time in the data prior to the outpatient or ED visit
- No diagnosis of influenza, influenza-like symptoms, or pneumonia in the preceding 60 days (to ensure the index date is the date of the most recent symptom of COVID-19)

The index date was defined as the date of the valid outpatient or ED visit.

For the external validation using non–COVID-19 data (influenza data), patients were included in the target population if they satisfied the criteria below:

- Presenting at an outpatient or ED visit with a record of influenza or influenza-like symptoms (ie, fever and either cough, shortness of breath, myalgia, malaise, or fatigue)

- Aged ≥18 years during the outpatient or ED visit
- ≥365 days of observation time in the data prior to the outpatient/ED visit
- No diagnosis of influenza, influenza-like symptoms, or pneumonia in the preceding 60 days (to ensure the index date is the date of the most recent symptom of influenza)

The index date was defined the date of the valid outpatient or ED visit.

Outcome

The outcome was hospitalization with pneumonia on the index date (valid outpatient or ED visit) and within 30 days after index.

Multimedia Appendix 2 contains the definitions of pneumonia, influenza, influenza-like symptoms, and COVID-19 used in this study. The full details of the participant cohorts and the outcomes used for validation can be found in the study package [13].

Predictors

The predictors of the logistic regression version of the C-19 index are age in years, male sex, number of inpatient visits during the prior 12 months, and indicator variables for various Clinical Classifications Software Refined (CCSR) categories. A table with the C-19 predictors and coefficients is presented in Multimedia Appendix 3. The CCSR categories used were pneumonia except that caused by tuberculosis, other and ill-defined heart disease, heart failure, acute rheumatic heart disease, coronary atherosclerosis and other heart disease, pulmonary heart disease, chronic rheumatic heart disease, diabetes mellitus with complication, diabetes mellitus without complication, chronic obstructive pulmonary disease and bronchiectasis, and other specified and unspecified lower respiratory disease. Age interactions with each CCSR variable were also included as predictors. Each CCSR category corresponds to an aggregation of International Classification of Disease, Tenth Revision (ICD-10) codes that belong to the category.

In the development data, if a patient had an ICD-10 code that was part of the CCSR "pneumonia except that caused by tuberculosis" grouping during a specified time period prior to index, their value for the predictor "pneumonia except that caused by tuberculosis" was 1; otherwise, it was 0. This assignment was repeated for each CCSR predictor. Data in the OMOP-CDM do not use ICD-10 codes, but instead use Systematized Nomenclature of Medicine (SNOMED) codes. Therefore, to replicate the predictors in the OMOP-CDM data, we needed to find the sets of SNOMED codes that corresponded to each CCSR predictor. We accomplished this by finding the SNOMED equivalent of each ICD-10 code in a CCSR category.

The SNOMED groupings per CCSR category used by the OHDSI implementation of the C-19 are presented in Multimedia Appendix 3.

Sample Size

We identified 41,381 patients with an outpatient or ED visit for COVID-19 in 2020: 1985 patients from South Korea, 37,950 patients from Spain, and 1446 patients from the United States. We also identified a total of 9,429,285 patients with an outpatient or ED visit for influenza or influenza-like symptoms in databases from six countries. The number of visits for influenza or influenza-like symptoms per database ranged between 2793 and 3,146,801.

Missing Data

The prediction models used a cohort design that included any patient who satisfied the inclusion criteria. We did not exclude patients who were lost to follow-up during the 30-day period after the valid outpatient or ED visit.

Statistical Analysis Methods

The model performance was evaluated using the standard discriminative metrics: area under the receiver operating characteristic (AUROC) curve (equivalent to the C statistic) and area under the precision recall curve (AUPRC). The latter is a useful addition to the AUROC when assessing rare outcomes [14]. An AUROC of 1 corresponds to a model that can always assign a higher risk to patients who will experience the outcome compared to those who will not. An AUROC of 0.5 corresponds to a model that randomly guesses a patient's risk. Precision is defined as the number of true positives over the number of true positives plus the number of false positives. Recall is defined as the number of true positives over the number of true positives plus the number of false negatives. The precision-recall curve shows the tradeoff between precision and recall for different thresholds. The AUPRC performance is relative to the rareness of the outcome. An AUPRC greater than the percentage of the population with the outcome indicates that the model is discriminating, and the greater the value (closer to 1), the better the discrimination. The AUPRC gives some insight into the false positive rate; a low AUPRC value indicates that the model will lead to many false positives. The calibration was determined by creating deciles based on the predicted risk and plotting the mean predicted risk versus the observed risk in each decile. If a model is well calibrated, the mean predicted risk will be approximately equal to the observed risk for each decile.

We followed the TRIPOD statement guidelines [11] for reporting the model validation throughout this paper. For transparency, an open source package for implementing the model on any OMOP-CDM data is available on GitHub [13].

Development Versus Validation

The differences between the C-19 index model development settings and the validation settings include a different target population and different data sets. Our validation design settings were chosen to mimic the situation in which a clinician needs to decide whether to admit a patient with COVID-19. Importantly, we validated the C-19 index model on patients with COVID-19.

The C-19 index was developed using a cohort design that entered adult patients into the cohort on September 30, 2016, and predicted whether they would be hospitalized for pneumonia or proxies (influenza, acute bronchitis, or other specified upper respiratory infections) in the following 3 months. Patients were required to have data for 6 or more months, and patients who left the database within 3 months of index and whose deaths

were not recorded were excluded. In our external validation, we used a cohort design but entered adult patients into the cohort when they had an initial outpatient/ED visit for influenza (or COVID-19) rather than a fixed date; also, we predicted hospitalization due to pneumonia in 30 days rather than 3 months. We excluded patients with influenza or pneumonia within the 60 days prior to index to restrict the data to initial visits. This mimics the situation during the COVID-19 pandemic in which clinicians need to decide whether to hospitalize a patient initially presenting with COVID-19. We required 12 months of prior observation and did not exclude patients who left the database within 3 months of index.

The C-19 index was developed using a subset of patients from the MDCR database prior to the pandemic. This is a US claims database containing patients aged 65 years or older. In this study, we were able to externally evaluate the C-19 index model on COVID-19 data, including adult patients under 65 years of age, from South Korea, Spain, and the United States.

Results

Web-Based Results

The complete results of our analysis are available as an interactive app [15].

The characteristics of the MDCR data (same data source as the development data but different patient subset) and the HIRA, SIDIAP, and VA data (patients with COVID-19) are displayed in Table 1. The characteristics for all data sets used in the study are available in Multimedia Appendix 4.

Table 1. Characteristics of patients at baseline in MDCR (database similar to the development data) and the data sets with COVID-19 data.

Predictor	Target population hospitalization during 30 days after index by data set							
	Medicare supplemental		HIRA[a]		SIDIAP[b]		VA[c]	
	Required	None	Required	None	Required	None	Required	None
Mean age (years)	80.92	76.41	65.53	45.09	63.28	49.61	69.64	58.07
Mean number of inpatient visits in prior 365 days	0.58	0.35	1.38	0.68	—[d]	—	0.32	0.22
Male sex (%)	52	45	56	46	59	43	95	80
Fraction of patients with a history of each condition (not including index)								
Acute rheumatic heart disease	0	0	0	0	—	—	—	—
Chronic obstructive pulmonary disease and bronchiectasis	0.43	0.25	0.38	0.21	0.06	0.03	0.27	0.21
Chronic rheumatic heart disease	0.03	0.02	0	0	—	—	—	—
Coronary atherosclerosis and other heart disease	0.19	0.15	0.21	0.09	0.02	0.01	0.17	0.13
Diabetes mellitus with complication	0.24	0.18	0.31	0.13	0.03	0.01	0.38	0.24
Diabetes mellitus without complication	0.38	0.32	0.43	0.20	0.13	0.05	0.50	0.32
Heart failure	0.37	0.20	0.20	0.07	0.02	0.01	0.23	0.12
Other and ill-defined heart disease	0.25	0.15	0.02	0.01	0.01	0.01	0.11	0.06
Other specified and unspecified lower respiratory disease	0.73	0.59	0.92	0.88	0.43	0.38	0.58	0.45
Pneumonia (except that caused by tuberculosis)	0.39	0.20	0.31	0.15	0.06	0.06	0.20	0.14
Pulmonary heart disease	0.09	0.04	0.00	0.00	—	—	—	—

[a]HIRA: Health Insurance Review and Assessment.

[b]SIDIAP: Information System for Research in Primary Care.

[c]VA: Department of Veterans Affairs.

[d]—: Data not included due to a low cell count.

Model Performance

When C-19 was transported to patients with COVID-19, it achieved AUROCs between 0.36 and 0.56; full details are provided in Table 2. The AUROC and calibration plots are presented in Figure 1. The internal discriminative performance of the C-19 index was an AUROC of 0.73. When we validated the model on patients in the MDCR database (patients aged ≥65 years with supplemental Medicare coverage), but with our target population consisting of symptomatic influenza patients, the performance was 0.65, a substantial drop from the development performance of 0.73. The AUROC performance when externally validated across other databases containing influenza patients ranged between 0.40 and 0.68. Full results are presented in Table 3, and the AUROC and calibration plots are presented in Multimedia Appendix 5. As a sensitivity analysis, we also validated the C-19 index on a target population consisting of patients who had COVID-19 or symptoms of the disease in 2020; the results were similar and are presented in Table S2 in Multimedia Appendix 1.

Table 2. External validation of the COVID-19 vulnerability index model on COVID-19 data. The target cohort was patients with an outpatient or emergency department visit with a COVID-19–positive record in 2020 and no symptoms in the prior 60 days.

Database	Target size, n	Outcome size, n (%)	AUROC[a] (95% CI)[b]	AUPRC[c]
HIRA[d]	1985	89 (4.48)	0.56 (0.488-0.636)	0.07
SIDIAP[e]	37950	1223 (3.22)	0.363	0.03
VA[f]	1446	149 (10.30)	0.529 (0.473-0.584)	0.14

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

[b]The 95% CI is reported when the outcome count is <1000.

[c]AUPRC: area under the precision recall curve.

[d]HIRA: Health Insurance Review and Assessment.

[e]SIDIAP: Information System for Research in Primary Care.

[f]VA: Department of Veterans Affairs.

Figure 1. Receiver operating characteristic and calibration plots of the COVID-19 vulnerability index model for the three data sets with sufficient and suitable COVID-19 data. HIRA: Health Insurance Review and Assessment; SIDIAP: Information System for Research in Primary Care; VA-OMOP: Department of Veterans Affairs– Observational Medical Outcomes Partnership.

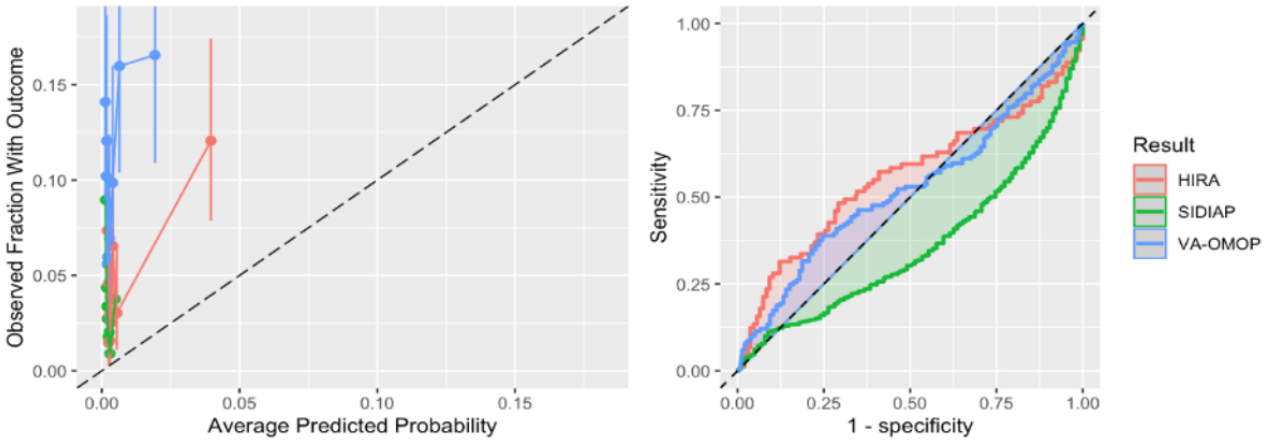


Table 3. External validation of the COVID-19 vulnerability index model on influenza patient data (non–COVID-19 data).

Database	Target population size, n	Outcome size, n (%)	AUROC[a] (95% CI)[b]	AUPRC[c]
Medicaid	536,806	32,987 (6.15)	0.68	0.16
Japanese Medical Data Center	1,276,478	728 (0.06)	0.58 (0.55-0.60)	0.004
Medicare supplemental	248,989	31,059 (12.47)	0.65	0.21
Commercial Claims and Encounters	3,146,801	33,824 (1.07)	0.58	0.04
Optum EHR[d]	1,654,157	34,229 (2.07)	0.62	0.07
ClinFormatics	2,082,277	105,030 (5.04)	0.67	0.17
Ajou University School of Medicine	3105	49 (1.58)	0.52 (0.41-0.63)	0.04
Tufts Medical Center Research Data Warehouse	6272	147 (2.34)	0.63 (0.58-0.69)	0.06
Australia Electronic Practice–Based Research Network	2793	29 (1.04)	0.59 (0.45-0.72)	0.03
Columbia University Irving Medical Center	27,356	1121 (5.10)	0.64	0.10
Integrated Primary Care Information	29,132	22 (0.08)	0.40 (0.26-0.54)	0.00
SIDIAP[e]	415,119	512 (0.12)	0.49 (0.45-0.52)	0.00

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

[b]The 95% CI is reported when the outcome count is <1000.

[c]AUPRC: area under the precision recall curve.

[d]EHR: electronic health record.

[e]SIDIAP: Information System for Research in Primary Care.

Discussion

The C-19 index is available on the web as a tool to predict severity in patients with COVID-19; however, it lacks validation for this population. Our validation across three data sets with sufficient COVID-19 data showed poor discriminative performance (AUROCs <0.6) and calibration. We observed similarly poor performance when the model was validated across 12 data sets with influenza patients, with the best AUROCs <0.70.

Interpretation

The key finding of this study is the performance of the C-19 index model when transported to patients with COVID-19. The model performance was poor (AUROCs 0.36-0.56) across the COVID-19 data sets. The performance was worse than random guessing in the SIDIAP data, which is consistent with the poor performance seen when the model was applied to European patients with influenza. The calibration plots show that the C-19 index consistently underestimated risk in the patients with COVID-19.

The data sets used to perform the validation had very different patient populations. MDCR had the oldest patient population, and many patients in this data set had comorbidities. Compared to MDCR, the CCAE and JMDC data sets presented healthier and younger patients (mean age approximately 40 years) in the target population. Although the MDCD data set contained younger patients, these patients often had comorbidities (ie, 20% these patients had chronic obstructive pulmonary disease, 11% had heart failure, and 17% had a history of pneumonia). The rate of hospitalization ranged greatly across the data sets, with values between 0.1% (JMDC) and 12.4% (MDCR). The rate of the outcome in the data set used to develop the C-19 index was 0.23%, much lower than that in the MDCR data set used to validate the model in this study. This is because our study was restricted to patients at the point they had an outpatient or ED visit due to influenza or COVID-19. Although five data sets contained patients with COVID-19, only four (VA, HIRA, SIDIAP, and CUIMC) contained sufficient data for external validation. The result of the C-19 index model when applied to patients with COVID-19 in CUIMC was poor, with an AUROC <0.5; however, this data set consisted mostly of hospitalized patients and therefore did not seem to be suitable for validating a model that predicts hospitalizations.

We chose a target population of symptomatic patients because this resembles the situation in which COVID-19 prediction models may be clinically implemented during the pandemic: clinicians would not be likely to admit asymptomatic patients. This suggests that the internal C-19 AUROC estimate, which was evaluated within the general population rather than among people with symptoms, may be optimistic compared to its use in a realistic setting due to the inclusion of many healthy patients in the model development data. When applied to predict hospitalization in influenza patients across US data, the discriminative performance ranged between 0.58 and 0.68**.** The performance was worse on the CCAE data set with younger patients, likely because age is a key predictor in the model. When the C-19 index was transported across non-US data sets, the discrimination was poor to reasonable in the Australian and Asian data (0.52-0.64) and poor in the European data (0.4-0.49). The European data were extracted from general practice settings, but the C-19 index model was developed using US claims data. Given the differences in clinical settings, it is not surprising that the performance was poor. This finding highlights that models often may not transport to different health care settings. The AUROC of 0.36 when the C-19 index model was validated in SIDIAP was worse than random guessing, and inverting the predicted risk would lead to an AUROC of 0.64. This may be a result of the C-19 including age interaction terms, which resulted in a negative age coefficient. Table 1 shows that in SIDIAP, the model's age-interacting comorbidities are not recorded as often as in the other databases. As a result, younger patients may have been assigned higher risks than older patients in SIDIAP.

The calibration was poor when applying the C-19 to COVID-19 data. This is not unexpected, as it is known that patients with COVID-19 have a higher risk of hospitalization due to pneumonia than the general COVID-19–free population. The calibration could likely be improved by performing recalibration using a sample of data from patients with COVID-19.

Implications

The results provide extensive insight into the performance of the logistic regression C-19 index when used for COVID-19 data. The external validation uncovered that the logistic regression C-19 index model is unreliable when predicting hospitalization risk for patients with COVID-19. Given this result, we do not recommend using the logistic regression C-19 index to aid decision-making during the COVID-19 pandemic. The model did not appear to transport to patients with COVID-19, highlighting the importance of externally validating models, especially models whose target population differs from the development population.

There are numerous potential reasons why the logistic regression C-19 index model failed to predict hospitalization due to pneumonia in the investigated patients with COVID-19. The first reason may be that the model was developed on patients aged 65 years or older but was applied to patients aged 18 or older. Age had a negative coefficient in the model, which may have caused issues when the model was applied to younger patients. A second reason may be due to incorrect phenotyping of the predictors. We matched the SNOMED codes to the CCSR ICD-10 codes provided; however, the predictors may require database-specific phenotypes due to coding differences across data sets and health care settings. This may explain the poor performance in the European data sets, which were obtained from databases that may record entries differently than those in the United States. A third reason is the study design. The C-19 index was developed to predict hospitalization from a set date in 2016; however, we validated it in a target cohort of symptomatic patients with an outpatient or ED visit, as this more closely matches the clinical use case of the model. Therefore, we were likely to have a sicker population, in which discrimination may have been more difficult. A fourth potential reason is that the C-19 index model was developed using data prior to 2017 but was validated on data from 2020: temporal

changes and concept drift may have negatively impacted the performance. Although we do not know the reason for the unreliability of the C-19 index model on patients with COVID-19, we were able to quantify it by large-scale external validation across a network of data sets. In future work, it would be beneficial to develop techniques that can identify reasons for poor external validation performance, as this may inform new best practices for model development.

This study highlights the importance of performing extensive external validation across different settings. During times of uncertainty, such as pandemics, medical staff who are under pressure to make important decisions could benefit from implementing vetted prediction models. However, it is important to gain an unbiased and reliable evaluation of a model's performance across numerous patient populations before the model is used. Internal validation can often be biased (eg, the population used to develop the model does not match the intended target population) and can provide optimistic performance estimates (eg, a poor design or small data set may result in overestimated discriminative performance). The approach used by the OHDSI collaboration enables efficient external validation of models across multiple data sets, and this is a valuable resource when urgency is required.

Limitations

A common issue when using observational health care data, especially across a network of databases, is the difficulty in developing phenotypes that are valid on all data sets. In this study, we used predictor definitions given by the researchers who developed the model. However, these definitions may not transport across all the data sets and may account for some of the decrease in performance. We were also limited to validating the less complex C-19 index model due to the large number of variables and lack of transparency for the more complex models.

The C-19 index model used in this paper to demonstrate the importance of external validation may have limited use for medical decision-making. Other COVID-19 models, such as those including physiological measurements, may be making more clinical impact. However, we choose the C-19 index model because it was available early in the pandemic and was being advertised to the public as a useful tool while being reported in a preprint paper with no formal peer review.

Conclusions

We have demonstrated the importance of implementing external validation in multiple data sets to determine the reliability of prediction models. We picked a newly developed model, the C-19 index, that aimed to predict which patients with COVID-19 are at risk of severe complications due to SARS-CoV-2. The model reported an internal AUC of 0.73 but was deemed to have a high risk of potential bias [2]. The C-19 index addresses an important issue that could have greatly aided decision-making during the COVID-19 pandemic; however, its performance in patients with COVID-19 was unknown. Our results show that the C-19 index performs poorly when applied to newly diagnosed patients with COVID-19 in Asia, Europe, and the United States. Overall, we suggest that the model currently only be used to predict hospitalization due to pneumonia in older patients in the United States. The results of this study demonstrate that internal validation performance should be considered optimistic estimates and that a prediction model requires validation across multiple data sets in the target population where it will be used (or a close proxy) before it should be trusted.

Acknowledgments

We would like to acknowledge the patients who have contracted or died of this devastating disease, as well as their families and caregivers. We would also like to thank the health care professionals involved in the management of COVID-19 during these challenging times, from primary care to intensive care units. The authors appreciate the health care professionals dedicated to treating patients with COVID-19 in Korea and the Ministry of Health and Welfare and the Health Insurance Review & Assessment Service of Korea for sharing invaluable national health insurance claims data in a prompt manner. This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This work was also supported by the Bio Industrial Strategic Technology Development Program (20001234, 20003883) funded by the Ministry of Trade, Industry & Energy (Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992]. This project is funded by the Health Department from the Generalitat de Catalunya with a grant for research projects on SARS-CoV-2 and COVID-19 disease organized by the Direcció General de Recerca i Innovació en Salut. The University of Oxford received a grant related to this work from the Bill & Melinda Gates Foundation (Investment ID INV-016201) and partial support from the UK National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. DPA is funded through a NIHR Senior Research Fellowship (Grant number SRF-2018-11-ST2-004). The views expressed in this publication are those of the author(s) and not necessarily those of the National Health Service, the National Institute for Health Research, the Department of Health, the Department of Veterans Affairs, or the United States Government. BSKH is funded through Innovation Fund Denmark (5153-00002B) and the Novo Nordisk Foundation (NNF14CC0001). This project is part funded by the University of New South Wales Research Infrastructure Scheme grant. SLD and MEM report funding from NIH NHBLI R-01, NIH NIDDK R-01 grant, and VA HSR&D. This work was supported using resources and facilities of the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI), VA HSR RES 13-457.

Conflicts of Interest

DPA reports grants and other funding from AMGEN, grants, nonfinancial support and other from UCB Biopharma, and grants from Les Laboratoires Servier outside the submitted work; also, Janssen, on behalf of IMI-funded EHDEN and EMIF consortiums and Synapse Management Partners, has supported training programs organized by DPA's department and open for external participants. PRR reports grants from Innovative Medicines Initiative and grants from Janssen Research and Development, during the conduct of the study. CGR and KK report that they are employees of IQVIA. JMR, PBR, AS, and GR are compensated employees of Janssen Research & Development, JNJ. MAS reports receiving grants from US National Institutes of Health, grants from IQVIA, personal fees from Janssen Research and Development, personal fees from Private Health Management, during the conduct of the study. DM is supported by a Wellcome Trust Clinical Research Development Fellowship (Grant 214588/Z/18/Z) and reports grants from the Chief Scientist Office, Health Data Research UK, and NIHR outside the submitted work. GH reports receiving grants from the US National Institutes of Health National Library of Medicine during the conduct of the study and from Janssen Research outside the submitted work. BSKH reports receiving grants from Innovation Fund Denmark and Novo Nordisk Foundation outside the submitted work. SLD reports grants from Anolinx LLC, Astellas Pharma Inc, AstraZeneca Pharmaceuticals LP, Boehringer Ingelheim International GmbH, Celgene Corporation, Eli Lilly and Company, Genentech Inc, Genomic Health Inc, Gilead Sciences Inc, GlaxoSmithKline PLC, Innocrin Pharmaceuticals Inc, Janssen Pharmaceuticals Inc, Kantar Health, Myriad Genetic Laboratories Inc, Novartis International AG, Parexel International Corporation through the University of Utah or Western Institute for Veteran Research outside the submitted work.

Multimedia Appendix 1
Supplemental tables describing the databases used in this study and sensitivity results.
[DOCX File , 18 KB - medinform_v9i4e21547_app1.docx]

Multimedia Appendix 2
Code set used to define each condition.
[XLSX File (Microsoft Excel File), 173 KB - medinform_v9i4e21547_app2.xlsx]

Multimedia Appendix 3
COVID-19 vulnerability index model and Systematized Nomenclature of Medicine phenotype codes.
[DOCX File , 19 KB - medinform_v9i4e21547_app3.docx]

Multimedia Appendix 4
Descriptive table for each of the non–COVID-19 case databases.
[XLSX File (Microsoft Excel File), 17 KB - medinform_v9i4e21547_app4.xlsx]

Multimedia Appendix 5
Receiver operating characteristic and calibration plots for the non–COVID-19 case databases.
[DOCX File , 1291 KB - medinform_v9i4e21547_app5.docx]

References

1. Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? Lancet 2020 Apr;395(10231):1225-1228. [doi: 10.1016/s0140-6736(20)30627-9]
2. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
3. Zhang H, Wang X, Fu Z, Luo M, Zhang Z, Zhang K, et al. Potential factors for prediction of disease severity of COVID-19 patients. medRxiv. Preprint posted online on March 23, 2020. [doi: 10.1101/2020.03.20.20039818]
4. Lu J, Hu S, Fan R, Liu Z, Yin X, Wang Q, et al. ACP Risk Grade: a simple mortality index for patients with confirmed or suspected severe acute respiratory syndrome coronavirus 2 disease (COVID-19) during the early stage of outbreak in Wuhan, China. SSRN Journal. Preprint posted online on February 28, 2020. [doi: 10.2139/ssrn.3543603]
5. DeCaprio D, Gartner J, McCall CJ, Burgess T, Garcia K, Kothari S, et al. Building a COVID-19 vulnerability index. ArXiv. Preprint posted online on March 16, 2020 [FREE Full text]
6. C-19 index. URL: http://c19survey.closedloop.ai/ [accessed 2021-03-05]
7. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376]
8. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc 2019 Dec 01;26(12):1651-1654 [FREE Full text] [doi: 10.1093/jamia/ocz130] [Medline: 31373357]

9. Hripcsak G, Duke J, Shah N, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
10. Reps J, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. Research Square. Preprint posted online on May 06, 2020. [doi: 10.21203/rs.2.11750/v3]
11. Collins G, Reitsma J, Altman D, Moons K. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg 2015 Feb 07;102(3):148-158. [doi: 10.1002/bjs.9736] [Medline: 25627261]
12. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
13. ohdsi-studies: Covid19 Prediction Studies. GitHub. URL: https://github.com/ohdsi-studies/Covid19PredictionStudies/tree/master/CovidVulnerabilityIndex [accessed 2021-03-05]
14. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
15. Can we trust the prediction model? Demonstrating the importance of external validation by investigating the COVID-19 Vulnerability (C-19) Index across an international network of observational healthcare datasets. C-19 Validation. URL: http://evidence.ohdsi.org/C19validation [accessed 2021-03-05]

Abbreviations

AU-ePBRN: Australia Electronic Practice–Based Research Network
AUPRC: area under the precision recall curve
AUROC: Area under the receiver operating characteristic curve
AUSOM: Ajou University School of Medicine
C-19: COVID-19 vulnerability
CCAE: Commercial Claims and Encounters
CCSR: Clinical Classifications Software Refined
CUIMC: Columbia University Irving Medical Center
ED: emergency department
EHDEN: European Health Data and Evidence Network
EMR: electronic medical record
HIRA: Health Insurance Review and Assessment
ICD-10: International Classification of Disease, Tenth Revision
IPCI: Integrated Primary Care Information
JU: Joint Undertaking
MDCD: Medicaid
MDCR: Medicare
NIHR: National Institute for Health Research
OHDSI: Observational Health Data Science and Informatics
OMOP-CDM: Observational Medical Outcomes Partnership Common Data Model
SIDIAP: Information System for Research in Primary Care
SNOMED: Systematized Nomenclature of Medicine
TRDW: Tufts Medical Center Research Data Warehouse
TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis
VA: Veterans Affairs

Edited by C Lovis; submitted 17.06.20; peer-reviewed by D Maslove, J Wang, A Austin; comments to author 28.10.20; revised version received 12.11.20; accepted 27.02.21; published 05.04.21.

Please cite as:
Reps JM, Kim C, Williams RD, Markus AF, Yang C, Duarte-Salles T, Falconer T, Jonnagaddala J, Williams A, Fernández-Bertolín S, DuVall SL, Kostka K, Rao G, Shoaibi A, Ostropolets A, Spotnitz ME, Zhang L, Casajust P, Steyerberg EW, Nyberg F, Kaas-Hansen BS, Choi YH, Morales D, Liaw ST, Abrahão MTF, Areia C, Matheny ME, Lynch KE, Aragón M, Park RW, Hripcsak G, Reich CG, Suchard MA, You SC, Ryan PB, Prieto-Alhambra D, Rijnbeek PR
Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study
JMIR Med Inform 2021;9(4):e21547
URL: https://medinform.jmir.org/2021/4/e21547
doi:10.2196/21547
PMID:33661754

Original Paper

A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth

Qian He[1*], BSc; Fei Du[1*], BSc; Lianne W L Simonse[1*], MSc, PhD

Department of Design Organisation & Strategy, Faculty of Industrial Design Engineering, Delft University of Technology, Delft, Netherlands

[*]all authors contributed equally

Corresponding Author:
Lianne W L Simonse, MSc, PhD
Department of Design Organisation & Strategy
Faculty of Industrial Design Engineering
Delft University of Technology
Landbergstraat 15
Delft, 2628CE
Netherlands
Phone: 31 15 27 ext 89054
Email: L.W.L.Simonse@tudelft.nl

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/5/e29794/

Abstract

Background: In the context of the COVID-19 outbreak, 80% of the persons who are infected have mild symptoms and are required to self-recover at home. They have a strong demand for remote health care that, despite the great potential of artificial intelligence (AI), is not met by the current services of eHealth. Understanding the real needs of these persons is lacking.

Objective: The aim of this paper is to contribute a fine-grained understanding of the home isolation experience of persons with mild COVID-19 symptoms to enhance AI in eHealth services.

Methods: A design research method with a qualitative approach was used to map the patient journey. Data on the home isolation experiences of persons with mild COVID-19 symptoms was collected from the top-viewed personal video stories on YouTube and their comment threads. For the analysis, this data was transcribed, coded, and mapped into the patient journey map.

Results: The key findings on the home isolation experience of persons with mild COVID-19 symptoms concerned (1) an awareness period before testing positive, (2) less typical and more personal symptoms, (3) a negative mood experience curve, (5) inadequate home health care service support for patients, and (6) benefits and drawbacks of social media support.

Conclusions: The design of the patient journey map and underlying insights on the home isolation experience of persons with mild COVID-19 symptoms serves health and information technology professionals in more effectively applying AI technology into eHealth services, for which three main service concepts are proposed: (1) trustworthy public health information to relieve stress, (2) personal COVID-19 health monitoring, and (3) community support.

(JMIR Med Inform 2021;9(4):e23238) doi:10.2196/23238

KEYWORDS

COVID-19; design; eHealth; artificial intelligence; service design; patient journey map; user-centered design; digital service solutions in health; home isolation; AI; touchpoint

Introduction

COVID-19

In December 2019, a new type of coronavirus causing acute respiratory syndrome (COVID-19) was discovered in Wuhan. COVID-19 spread rapidly around the world and was designated as a Public Health Emergency of International Concern by the World Health Organization (WHO) on January 30, 2020 [1]. By 2 AM CEST (Central European Summer Time) on June 29, 2020, there had been 9,962,193 confirmed cases and 498,723 confirmed deaths [2]. Although lockdown measures have been eased in many countries, a WHO COVID-19 situation report (June 28) still shows a rising trend [3]. This worldwide pandemic has had a far greater impact than expected, and the upward trend is likely to continue in the near future until effective vaccines and antivirals are introduced.

Home Isolation for 80% of Persons With Mild COVID-19 Symptoms

Looking back at the early months of the outbreak, the rapid spread of COVID-19 and the growing number of patients placed a burden on unprepared medical systems worldwide [4,5]. Some countries such as China and Spain set up mobile cabin hospitals to relieve the pressure on their hospitals [6-8]. However, to ensure that the limited health care resources were spent on urgent cases involving severe symptoms, most countries decided that persons with mild symptoms could be isolated at home for a self-recovery trajectory [9]. According to the early WHO studies, about 80% of COVID-19 cases present mild symptoms, and most of these patients should typically be able to recover at home [10-13]. Therefore, disease control centers in various countries (eg, the United Kingdom, the United States, the Netherlands, Italy, and Canada) directed persons with mild COVID-19 symptoms to stay at home and contact their general practitioner by phone instead of directly visiting the hospital. New guidelines for home care were developed by the WHO and countries' public health departments to present the proper measures for the home care of patients [13-17].

Strong Demand for eHealth Services

As an alternative solution to conventional health care services, the uptake of eHealth services rapidly expanded during the COVID-19 pandemic [18]; the main fields of application have been telemedicine, remote patient monitoring, and triage and risk assessment [19]. Enabling better response to the pandemic, such digital health care solutions not only reduce the risk of disease transmission thanks to the provision of remote medical care services but also hold the promise to improve the mental health of isolated patients with distance guidance [18-20]. Initial studies have shown that, since the outbreak began, the frequency of internet searches related to "online medical care" has increased significantly, and the public's interest in eHealth is rising as the number of infected cases climbs [21,22]. Although the limited capacity of the existing eHealth service systems cannot immediately meet this growing demand [22], there is no doubt that the COVID-19 pandemic will impact the current service provision of medical institutions and lead to an accelerated transition to digital health care.

Potential of Artificial Intelligence

The large number of digital health applications that have been released in response to the COVID-19 outbreak includes a growing number of artificial intelligence (AI) tools; these include tools that make use of natural language data processing and machine learning with big data lakes, such as in decision support agents, advanced self-diagnoses tooling, and AI-enabled mental health interventions [23-27]. These AI tools have the potential to add new service options to remote health care modes such as remote assessment, remote diagnosis, remote interaction, and remote monitoring [24]. The main touchpoints of AI technology appeared to be mobile phone apps, wearable devices, and chat robots [25]. Prior research in the context of public eHealth and disease prevention has found that AI technology improves patients' health conditions more efficiently [24]. Moreover, application of AI appeared to enable more personalized care pathways based on personal health profiles [18]. The COVID-19 pandemic requires existing health care models to have better integration, delivery, and distribution capabilities, and thus, new requirements for AI in eHealth have been put forward [18,28,29]. Thus far, the main fields that AI technology has been applied to during the pandemic is early detection and diagnosis of infections, personal contact tracking, case and mortality prediction, drug and vaccine development, reducing the workload of medical staff, and other aspects of controlling and managing the spread of the virus [23]. Overall, the emerging AI tools have the potential to perform a useful role in combating COVID-19, and in particular, AI has the potential to improve the quality of home care by providing more personalized, sophisticated, and continuous medical services [18,24].

A Lack of Understanding of the Real Needs: Home Isolation Experiences

Despite the potential of AI in digital health services, many attempts to integrate AI technology into health have failed [24,30]. The primary reason for this is that the development of AI tools and applications is predominantly focused on technical and functional aspects, and largely ignores the demands of users and contextual aspects [31]. Data scientists and health professionals usually start developing AI technology without exploring the patient perspective in advance, that is, the health experience and needs of the users. Often users are only involved in providing feedback after a system solution test is carried out or after the final digital service has been released [30]. However, without sufficient understanding of the user experience, AI service applications will not be capable of solving the real and serious problem of a lack of useful value [32,33]. In other words, the real needs of patients with mild symptoms of COVID-19 have not yet been identified. Illustrative of this lack of understanding the real needs, most of the current machine learning algorithms behind decision support tools are too opaque and difficult for users to reconstruct [34]. When AI machines provide treatment suggestions, the mysteriousness of this process basically causes users, including health professionals and patients, to question the result [35]. If AI technology is expected to be used to improve existing eHealth service capabilities, the key actors should focus more on the users rather than technology to reduce the gap between technology and user, and to improve

the usefulness of AI [24]. Corresponding findings related to cancer conditions indicated that AI technologies provide a way to transition from a traditional aperiodic "snapshot" monitoring approach to a continuous and longitudinal monitoring paradigm, increase patients' engagement in their care, and facilitate doctor-patient interaction pathways [36]. In particular, a study that applied machine learning and natural language processing techniques on social media data from online cancer support groups provided new insights toward informed decision making on personalized health care delivery [37]. Likewise, an increasing amount of AI demonstrators evidence a new service potential of AI applications that are hardly used yet in enhanced service providing. To be able to meet personal patients' needs, in-depth research is required.

Design for Better Supported Home Isolation Experiences

To overcome the barrier of the lack of knowledge on what is useful and what is not, AI research and development should involve user-centered design methods to gain insights into the real experiences and needs of users [38], as well as apply a person-centered perspective to construct an explainable AI and make the AI process more transparent and comprehensible to multiple users, including patients and health professionals [39]. As Xu [32] stated, "a useful AI is defined as an AI solution that can provide the functions required to satisfy target users' needs in the valid usage scenarios of their work and life," which means the user experience should be deeply understood before the AI development starts. However, most of the current research is mainly focused on persons with severe COVID-19 or patients treated at hospitals. We found no literature that explicitly describes the home isolation experience of persons with mild COVID-19 symptoms. Thus, knowledge of this area is still required. This paper intends to address this gap by elaborating on the entire process of the home isolation experience of people with mild COVID-19 symptoms—from infection to recovery—and then extracting in-depth insights for AI concepts in eHealth. Specifically, our research question is how can we improve the home isolation experience of persons with mild COVID-19 symptoms through eHealth services with AI technology?

Methods

Design Research

We employed a design research method in which a qualitative approach was used to explore the home isolation experience of persons with mild COVID-19 symptoms because this is a relatively new field [40]. Our design research had a phenomenology perspective that rests on the philosophical assumptions of studying people's experiences in their daily living, viewing these experiences as conscious [41]. This phenomenon study provides a real and comprehensive description of the home isolation experience of persons with mild COVID-19 symptoms, which is needed to obtain insights into their user needs and tasks [40], and to find touchpoint interaction needs for the useful application of AI in eHealth.

Patient Journey Mapping

Patient journey mapping is a method of design research for developing health care services from a patient perspective [42-45]. The purpose is to capture insights into a patient's activities, interactions, feelings, and motivations throughout the personal health care journey, and to generate insights into user values and dilemmas that lead to the identification of real and underlying problems that must be solved through the successful application of innovative solutions [42,45]. The final journey map visualizes the commonly shared patient experiences and includes both physical, rational, and functional aspects of the patient experience as well as the emotional, interactional, and feelings aspect of patient experiences [42]. The design quality of the patient journey map is determined by its properties to visualize the knowledge and insights about the patient's experience and enable sympathy of the viewers by placing them in the perspective of the patient [44].

Prior research exemplified concise journey maps of visually compelling stories, distilling research into all aspects of personal experience and informing the reflections on the steps and approach laid out in the patient journey method [42]. The design of the patient journey map in this study contributed new knowledge on the home isolation experience of persons with mild COVID-19 symptoms and thereby provided the view of the patient and enabled a deep understanding of their whole experience from the onset of illness to recovery [45]. The patient journey map depicts all steps of the journey to gain a better understanding of the whole journey, taking the arising needs of patients into account, and uncovers new research fields for relevant AI applications [46].

Data Collection

Data on the home isolation experiences of persons with mild COVID-19 symptoms was collected from the top viewed personal video stories on YouTube and their comment threads. As researchers, the global pandemic meant that we were bound by the necessity to engage in social distancing and limit interpersonal contacts. Therefore, we chose to use personal video stories instead of interview techniques. YouTube, one of the major online video sharing platforms, has become recognized as a valuable social media source for personal stories about health and disease [47]. We chose YouTube as the data source because, compared to other social media platforms such as Twitter, Instagram, and Tik Tok, its content richness and the completeness of the stories presented on it enables more detailed data collection about an entire journey experience. Videos also allowed us as researchers to better understand the feelings of the patients from their nonverbal movements and expressions [48]. The personal video stories and comment threads provided us with a new research opportunity to investigate actively shared experiences instead of relying on actively obtained experiences from interviews. From a researcher's perspective, YouTube videos eliminated research bias and brought to light unexpected information that those posting them consider important from their personal perspective. The difficulties involved in the use of YouTube concerned the analysis of different narrative structures that posed challenges in extracting, coding, and synthesizing commonly shared experiences.

Sample Strategy

Purposive sampling was selected for the in-depth study of the experiences of persons with mild symptoms who were in home isolation [49]. As most YouTubers are young people, who account for a large proportion of persons with mild symptoms, they do not represent the total population of persons with mild COVID-19 symptoms across socioeconomic classes and ethnic and cultural groups [47]. We were likely to find personal video stories on home isolation experiences from the young YouTubers population representation because they are more likely to become the first embracer of new technology-based services [50]. The YouTubers we selected for the study after using the search terms "COVID-19," "experience/personal stories," and "home isolation" were, first, persons who shared their COVID-19 health conditions over a period of consecutive days or weeks. Second, we looked for influential videos with more than 100,000 views. Third, we selected stories that were perceived to be authentic and did not have any negative comment about their authenticity from more than 100,000 views, and we excluded those videos that did. In addition, we checked if the content stayed available (dated December 16, 2020) after the YouTube COVID-19 Policy and Security. Fourth, to achieve data saturation, we selected 5 as a suitable sample size to cover the wide range of possible experiences [51]. To some extent, this study is representative. In particular, the use of the YouTube platform comes with constraints for validity, as it can only represent the internet population and, in our sample of the videos with the most viewers, those who use the English language and live in the region of the Americas and Europe [47,48]. Further constraints relate to the fact that upper middle class Americans of European decent are more likely to post [47]. Table 1 lists the characteristics of the sampled personal video stories. To dig further into how online social support influences persons with mild COVID-19, we also collected and analyzed the top 50 popular comments on each of the videos.

Table 1. Sample of personal video stories and comment threads.[a]

No.	YouTuber's age (years)	Mild COVID-19 health and well-being condition (when uploaded)	Language of personal video story	YouTuber's region	Upload date in 2020	Video length (MM:SS)	Views, n	Likes, n	Dislikes, n	Comments, n
1	20-30	Sick	English	Americas	March 10	11:48	661,846	58,000	935	7082
2	20-30	Almost recovered	English	Americas	March 25	10:03 + 13:37	395,069 + 61,749	6677 + 1629	567 + 50	2350 + 725
3	20-30	Almost recovered	English	Europe	April 11	43:15	248,659	5637	333	1208
4	20-30	Almost recovered	English	Americas	April 9	11:27	183,711	3527	254	1315
5	41	Sick	English	Europe	April 5	10:50	246,814	15,000	263	3315

[a]Data collected from YouTube on May 26, 2020 (checked on December 16).

Ethics

This study was reviewed by the Human Research Ethics Committee of the Delft University of Technology [52]. In our sample strategy, we did not involve vulnerable groups of children or patients older than 65 years. As the personal video stories are published on the YouTube platform, we considered that they, in principle, constitute a publicly available data source for research [53]. For further confirmation, we emailed all 5 YouTubers to obtain consent and received 2 replies with affirmative answers. To minimize potential harm, we kept their identity anonymous and did not describe their characteristics and contexts in detail.

Data Analysis

In the data analysis, triangulation was used by clustering the data from the observation and the transcripts of videos and the comment threads [54].

Patient Journey Mapping

To fully understand the experience of these patients during home isolation, both generic and personalized experiences were analyzed based on the similarities and differences between the patients, respectively [42]. The analysis of the indicated *stage duration* was visually mapped to make the similarities and differences between personal journeys transparent. The *symptoms* of each patient at different stages were analyzed and mapped (see Multimedia Appendix 1).

From the transcript and narrative structure of each personal video story, quotes about their *doing, feeling, and thinking* were extracted and initially mapped separately into 4 personal journey maps. The *stages* were framed and labeled based on the similarities of activities and interactions across the first 4 personal journey maps. We generated the commonly shared journey map and added one more personal video story (the fifth YouTuber video), extracted the quotes, and analyzed the activities and interactions to check whether we had reached theoretical saturation on the generated journey map (see Multimedia Appendix 2).

Commonly mentioned symptoms come first. We then detailed the *steps* within each stage based on the "doing" quotes in transcripts. Based on the combination of data on feelings, steps, and symptoms, the *mood experience curve* was created to clarify the generic mood experience of the patients during the whole journey, from when they became aware of incipient symptoms to quarantine and self-recovery. Since not all of these persons with mild COVID-19 went through all steps, we bolded the video timeline to indicate which steps each patient actually experienced. The *video timeline analysis* of the video duration of each stage per patient was mapped with the percentage (divided by the total video duration), indicating which stages

the patient attached more importance to (see Multimedia Appendix 1). Finally, from the analysis of the *touchpoint interactions*, the specific services and products were clustered, categorized, and mapped on the resulting patient journey map (visualized in Multimedia Appendix 2).

Comments Thread Analysis

The YouTubers usually mentioned the purpose of publishing the video at the beginning or the end of the video, and the comments were responses to the YouTubers. From the more than 1000 comments per video, we selected those 50 comments that had interaction in the form of a follow-up comment from the YouTuber.

We thus analyzed the *video purpose* together with the comment threads to figure out the interaction between the YouTuber and the viewers, and combined these data sources to analyze the underlying purpose for sharing the home isolation experience in depth. In the transcripts, we annotated the quotes concerning why they wanted to post the video by means of initial coding, then classified the purposes mentioned by different YouTubers and synthesized them into a classification [55] of 4 themes in Multimedia Appendix 3.

Since each comment expressed more than one meaning and there were overlaps between different comments, we used an Excel (Microsoft Corporation) table to code the *comment threads*. First, we put the top 50 comments on each video in the first column in an Excel table, put the initial codes in the first row, and marked the cells where they intersected. Second, we counted how many viewers mentioned each code. Finally, we categorized the codes into 13 themes (see Multimedia Appendix 3) and pointed out how many people mentioned each theme in the 250 comments.

Touchpoint Needs Analysis in Relation to AI in eHealth Services

Based on this data analysis of the patient journey map and interaction between YouTubers and viewers, we synthesized key insights by inductive reasoning [56] and clustered key insights into 13 categories, leading to three identified needs of persons throughout the journey of home isolation (see Multimedia Appendix 4).

Results

Key Findings

The patient journey in Multimedia Appendix 2 maps the commonly shared home isolation experiences of persons with mild COVID-19 symptoms. The first key findings concerned an extensive awareness period before testing positive, experiences of less typical and more personal symptoms, a severe negative mood experience curve, and inadequate home health care service support for patients with mild COVID-19 through all stages. Second, the key finding from the analysis of the video's purpose and the comment threads concerned the benefits and drawbacks of social media support for patients with mild COVID-19. With the third and final key findings, main touchpoint needs during home isolation were synthesized to provide opportunities for AI eHealth concepts.

Awareness Period Before Testing Positive

The stories on personal experiences revealed a considerable period during which the YouTubers became aware of the outbreak of the virus and its public health impact before they related their symptoms to COVID-19. Although most of these persons (P2, P3, and P4) became highly aware of the public health threat (after the prestage of unconsciousness) in less than a week, some of them had low awareness (P1 and P5) and took much longer to do so—from 2 weeks to as long as 2 and a half months (stage 1). All of them went through a period of up to 4 days during which they related the public health situation to their personal condition and symptoms (stage 2), followed by 1-2 days for the testing stage (stage 3). The home isolation period (stage 4) lasted at least 1 and a half weeks but was around 1 month for most (P1 and P2). As none of the YouTubers had yet fully recovered at the conclusion of their video stories, the self-recovery period is expected to last even longer.

Less Typical and More Personal Symptoms

Based on the overall analysis of similarities and differences, the symptoms reported in the personal stories appeared to be different from one another. Each of the patients appeared to experience their own specific symptoms. All in all, almost 50 different symptoms were reported, ranging from a mild headache, loss of smell, a stomachache, high temperature, and dizziness to the more critical symptoms of fainting, shortness of breath, and high heart rate (Multimedia Appendix 1, bottom layer). In addition, the occurrence of similar symptoms also appeared to be different over time. For instance, P1 and P2 only had a fever in stage 2, while P4 had a continuous fever until the end. In contrast to these individually experienced physical symptoms, a general consensus was found on negative feelings and deteriorating mood experiences.

Negative Mood Experience Curve

The consensus on the negative feelings that all the YouTubers experienced concerned severe anxiety about dying and related feelings of depression and despair.

> *I knew I was gonna get sick and we'd go through the process that we're seeing on TV, go to the hospital, have complications and die. This is horrible to think about. It's so, so scary* [P2]

From the moment that they experienced their first symptom, they experienced severe negative moods that became dramatically worse when the symptoms continued to deteriorate, reaching the lowest level just before testing positive. (The mood experience curve is diagrammed on the top layer of the patient journey in Multimedia Appendix 2.) Surprisingly, testing positive was commonly experienced as an emotional relief. After this, their overall negative mood improved slowly but gradually during the home isolation period of self-recovery. That said, some of them experienced another period in which they felt emotionally broken again and then improved afterward. It is worth noting that, although not all of the patients went through the same ups and downs, overall, they all faced severe feelings of depression and mood fluctuations throughout the journey and especially in stages 2, 3, and 4.

The main differences were that P3 and P5 did not repeatedly look for medical help with no improvement in the second stage while P1 and P2 did. P3 was the only one who had not been tested and did not worry about having limited access to resources because her sister is a doctor and can get timely 24-hour professional help.

Inadequate Home Health Care Service Support

As shown in Multimedia Appendix 2, the patient journey followed several distinct stages: prestage with unconscious and low awareness of the public health risk posed by the virus outbreak, experiencing suspected symptoms, relating symptoms to COVID-19, testing and confirmed positive, and quarantine and self-recovery.

Prestage: Unconsciousness

The patient journey starts in the stage in which the patient is still unaware of the situation (prestage of unconsciousness). This is the stage that the patients tried to recall to reconstruct how they were infected. All had been to public areas and crowded places such as supermarkets, cafés, and party locations.

> *I was at a party at one of the hotels – there are probably over a thousand people* [P1]

Unaware of the spread of the virus and the danger of becoming infected, most of them also continued to visit public places.

> *The thought that I could have been infecting other people is just horrific to me* [P5]

This situation caused particular feelings of guilt about their personal and public responsibility for having infected others before being diagnosed with COVID-19 (P3, P4, and P5).

First Stage: Experiencing Suspected Symptoms

At the beginning of the first stage, before experiencing any symptoms, some of the patients already became anxious about the news of the COVID-19 outbreak. When the first symptoms were appearing, the mood of most of the patients began to decline rapidly, worsening as they experienced more physical symptoms and paid more attention to media coverage of the abnormality of the hospital situation.

> *Before my family got sick, my anxiety about all this was pretty high* [P2]

Those who were highly aware of COVID-19 could quickly relate their own symptoms to COVID-19. Others mentioned they had little knowledge about COVID-19 until the moment when they got tested and diagnosed positive.

> *I'd been sick for two months and I still did not have an answer, I still had all the symptoms* [P1]

This had major consequences, as they had not taken enough appropriate protective measures and infected several others—the longest period a person went without diagnosis was 2 and a half months.

Second Stage: Relating Symptoms to COVID-19

In the second stage, when the YouTubers started to realize that they had a high possibility of being infected, some could accept it, while others could not.

> *It's not corona, I think it's laryngitis, fingers crossed* [P3]

Most became highly anxious and even panicked due to the overwhelming media coverage and "death statistics" on patients with severe COVID-19 in hospitals. They started to have dark thoughts and different levels of stress up to severe depression.

> *Since I got sick who would a guessed, wrote down a few notes cuz my mind is like scrambled* [P2]

These persons indicated that they paid too much attention to COVID-19, and the overwhelming negative information led them to live in constant anxiety. In addition, all of these persons experienced a lack of medical help and guidance. Due to this lack of help, some chose to endure all the symptoms to save medical resources for others.

> *I didn't necessarily want to go to the emergency room because I didn't want to take resources away from people who needed it* [P4]

The only positive spark that provided a little comfort was the help they received from their family and friends.

> *Then all of a sudden I just fainted so I got up and I tried to go get my roommate in case anything happened* [P4]

Overall, due to the inadequate and ineffective support from professional health care, most of these people constantly worried about COVID-19 and its terrible consequences. All reached the lowest level of severely negative mood at the end of this stage.

Third Stage: Testing and Confirmed Positive

In the third stage, when the persons began to seek clinical support to test their suspicion that they had the COVID-19 virus themselves, most experienced an improvement in their mood. However, some of the others with mild symptoms were not diagnosed with COVID-19 at the first consultation due to a lack of clinical knowledge about mild COVID-19.

> *I did all the tests and he could not figure it out, now the one thing he did know was I was still having night sweats* [P1]

These persons became severely upset about the ineffective treatment they personally experienced, and their videos provided examples of incorrect diagnoses and repeated visits to health professionals.

> *He did a bunch of tests and they all came back negative. They didn't test me for COVID-19 though because they just said that they had to keep that for people who really needed it* [P4]

Fourth Stage: Quarantine and Self-recovering

In the fourth stage of home isolation and self-healing, the mood of all the patients tended to fluctuate several times, as they refrained from social interaction for a long time and had unstable health conditions. They still felt depressed when their health deteriorated again during isolation.

> *It's definitely the sickest I've ever been in my adult life* [P5]

They required professional guidance on the proper measures to take while in quarantine at home because they were concerned about infecting family members.

> *We talked to the doctors and they said like there was no reason that I had to stay separated from everybody (whole family infected)* [P2]

The main reasons for an improvement in mood were that the symptoms had been mild and were getting better, and the patients were taking on more activities and gradually returning to a normal life. The reasons for feeling more negative were the abnormality of life in home isolation, new severe symptoms, and ineffective treatment. Overall, the mood of these patients improved particularly after they received social support and effective treatment.

Benefits and Drawbacks of Social Media Support

The findings from the analysis of the YouTubers' purposes for sharing their videos and the comments made by viewers confirmed the benefits and drawbacks of the social sharing of public health experiences. The video purpose analysis revealed the commonly shared purpose of going through a difficult time together and receiving support from the audience. Reasons for sharing their personal story were to encourage viewers to pay more attention to protective measures and take social distancing seriously in public; to help viewers relieve their excessive anxiety and fear, and gain a better understanding of mild COVID-19; and to help others with similar experiences by sharing their real condition and self-recovery advice.

The findings from the comment analysis showed that the majority of the 250 commenters (n=166, 66.4%) expressed their likes and thanks to the YouTubers for sharing real COVID-19 experiences and encouraged and blessed them. Of the commenters, 25.2% (n=63) also shared their experiences and feelings, indicating that they can relate to the YouTubers. After watching the video, 13 of them said that they actually realized that they might have mild COVID-19 too. Some (n=26, 10.4%) indicated that they became scared and depressed. A minority (n=48, 19.2%) talked about the public health response to COVID-19 from governments, media, and the public, and asked people to take it more seriously. Home remedies such as vitamin C, elderberry syrup, lemons, and honey were suggested by 7.2% (n=18) of the commenters. Only 5 health care professionals commented. Inappropriate behavior such as going for a walk before having fully recovered were pointed out by 9.2% (n=23) of the commenters, and some (n=10, 4%) of the commenters made jokes.

In summary, the positive influence of personal video stories is that they reach people who are not familiar with the disease yet, they encourage viewers to take mild COVID-19 more seriously, and they provide some emotional relief; their negative influence is that they can spread disinformation and panic.

Main Touchpoint Needs During Home Isolation

Based on these key insights of the patient journey map and interaction between YouTubers and viewers, three needs were identified.

Touchpoint Need 1: Stress Release

Concerning the touchpoint interactions, the patients commonly mentioned the difficulty of obtaining trustworthy information. Although information about COVID-19 was easily available from various sources such as TV, friends, and websites, the quality of these sources varies.

> *I think someone sent me yesterday an article with no one delaying conditions dying but it's still kind of like really freaks you out when you're home and can't breathe* [P3]

The patients found it hard to judge the truthfulness of news. An incorrect perception of the disease resulted in continued aggravation of the symptoms and brought a strong sense of uneasiness and anxiety to the patients with COVID-19.

When symptoms first appeared, the YouTubers wanted to find out the cause of their physical discomfort. Due to their lack of knowledge of all the COVID-19 symptoms at the beginning of the outbreak, many of them behaved as they would with a normal disease. However, their continuous uncertainty, ineffective treatment, and deteriorating condition caused them fear and anxiety. Care professionals working in regular health care services were not able to diagnose patients with mild COVID-19 with atypical symptoms, which led to a long period of uncertainty (the longest of which was 2.5 months).

Touchpoint Need 2: Personal Health

Patients with mild COVID-19 had a need for professional medical guidance throughout the journey, beginning from when the symptoms appeared, with a focus on different needs at different stages.

> *Everyone's kind of dealing with like some symptoms but there's no confirmation because they couldn't give us the testing, so this is kind of where we are at until this next super weird symptom hit* [P2]

The strong feelings of uncertainty and stress due to negative thoughts caused an urgent need for testing; rejection could increase the negative impact on mental health. After testing positive, some of the patients became excited when their physical condition temporarily improved and then experienced mental breakdowns when their condition became worse again.

> *I just isolated at home and [did] not go out at all until three days after all the symptoms disappear[ed]* [P2]

It is hard for people to judge the point of recovery without a professional diagnosis.

Touchpoint Need 3: Social Support

In all their personal video stories, the YouTubers mentioned that their families, friends, and viewers provided them with plenty of help, ranging from basic support for living to emotional support for coping with anxiety. When they first felt a strong sense of insecurity due to the onset of weird symptoms, they longed for help from their families and friends to obtain basic necessities like food and medicine.

> *I find it hard to do little things like clean my teeth then go for a shower. I couldn't get my hands on any*

> *paracetamol for weeks. I really don't know what I would have done without it* [P3]

From the moment they suspected they were infected and went into home isolation, their internet communication became more important.

> *I'm separated from my family, I can't see my son or my wife* [P5]
>
> *I am on the phone with my friend and Facetime regularly* [P3]

Most of them could no longer work, study, or engage in their usual hobbies, and they experienced feelings of boredom and frustration about this, although some started to enjoy a new hobby. Overall, all of these people felt lonely and helpless during isolation. They shared their story on YouTube with the aim of helping others who had COVID-19.

Discussion

Principal Results

This study clarifies the stages, symptoms, mood curve, and touchpoint needs of persons with mild COVID-19 symptoms through mapping the patient journey. This design research of the systematic and in-depth analysis of how patients with mild COVID-19 told their personal stories in their self-shared videos and the comment threads found that persons with mild COVID-19 usually took an extensive period to realize that they were personally experiencing the public health threat of the virus outbreak. They all faced the same problems of severe negative and fluctuating moods while dealing with different symptoms. They lacked adequate and effective home health care service to overcome adversity. The home-isolated persons with mild COVID-19 symptoms turned to their family and friends not only for social support but also for medical assistance and obtained additional emotional support by sharing their stories on social media. Three principal touchpoint needs were identified. First, there is a need to relieve the anxiety caused by the virus by providing reliable public health information. Second, more personal health monitoring and guidance is needed to address personal symptoms. Third, more mental health guidance and social support is required to positively influence the severe moods and emotional problems of those with mild COVID-19.

The theoretical implications concern a new contribution to better understand underserved persons with mild COVID-19 symptoms during their home isolation. As a contribution to the field of AI in eHealth, we propose taking the user-centered findings and embedding them in AI eHealth service touchpoints to improve the home isolation experience.

Proposed Service Touchpoints of AI

As the number of patients with COVID-19 is still increasing and some countries are still conducting limited testing, the shortage of global medical resources will persist in the near future, and the demand for more and better eHealth services for patients will continue to rise. To meet the urgent need for public health, it is time to put AI technologies into practice. Although persons who are unfamiliar with new technologies may be less willing to use them, research also shows that, as long as they feel that a specific eHealth service has the ability to improve the quality of treatment, they will intend to try it. Therefore, the target users for eHealth innovation with AI are those who have urgent needs for better medical service—in this study, the target was home-isolated patients with mild COVID-19 symptoms in the context of limited medical resources. In addition, as most of those with mild COVID-19 symptoms are young people, the application of AI can be easier to promote because they are generally more receptive to new technologies [50].

We translated the patient journey insights into value creation for AI innovations in eHealth [43] and designed 3 initial service concepts for an AI application. We used the insights to improve a patient's experience with eHealth services. The three concepts are based on the premise of an eHealth app used on smart mobile devices.

Trustworthy Public Health Information

In this concept, persons can get more trustworthy information about COVID-19.

- *Group identification*: Identify the group of people who do not pay attention to the outbreak through social media and automatically show more information about COVID-19 in the areas of interest they often follow.
- *Dangerous area identification*: Identify dangerous areas by tracking patients who were diagnosed and public transportation data. Evaluate the hazard level. Release information about dangerous areas to remind people that they should visit these areas less often and take protective measures while paying more attention to their physical condition if they have been in dangerous areas.
- *Symptom analysis*: Collect all the atypical symptoms related to COVID-19 that are shared on the internet, reminding the public to pay attention to these symptoms. Meanwhile, facilitate the work of health professionals to better study COVID-19.
- *Information analysis*: Based on the user's search history, provide more information on the issues that cause the most anxiety to the user and that they have the most questions about. Meanwhile, have experts identify and refute false information or rumors.
- *Positive relaxation*: Provide personalized information for users who allow the use of their data. Show more related information in line with their interests. If the user is experiencing depression because of COVID-19, present more positive information and stories of patients with mild symptoms who have recovered.

Personal COVID-19 Health Monitoring

In this concept, by inputting symptoms and physical condition data through text or voice, users can self-diagnose whether they have been infected by COVID-19 and follow up with self-monitoring and personalized care as well as daily predictions about their potential health condition.

- *Self-check*: AI carries out a preliminary diagnosis based on the symptoms indicated by the user and answers to questions, and provides a diagnosis result in the form of a list of all the possible causes with their probability as a

percentage, especially the possibility of being infected with COVID-19.

- *Remote diagnosis*: According to the user's recorded symptoms and physical data, the system can automatically match a suitable general practitioner or specialist for the user to communicate with while making it easier for the doctor to arrive at a diagnosis and give treatment recommendations.
- *Controllable testing process*: Based on the user's health condition records, AI prioritizes users with severe symptoms (this process runs in the background to prevent users from recording false information because they want a test as soon as possible). AI recommends the most suitable hospital and shows the potential waiting time. All medical assistance provides a clear status report on progress and the estimated waiting time for the results to improve patients' feeling of control.
- *Professional advice*: According to the diagnosis result, the system gives suggestions for the next steps. If the probability of infection is low, the system will suggest that the user should continue paying attention to their physical condition and take proper protective measures when going out. If the probability is high, the system will suggest that the user should go into quarantine and continuously observe the symptoms for a few more days. In addition, based on the user's symptoms, the system will present similar cases to help users better understand the disease.
- *Self-monitoring and treatment*: Users can connect their monitoring device such as an oximeter to the app to automatically collect body data or manually record their body condition daily. The system judges the development of the disease daily based on the data. It also provides proper treatment according to the user's health condition (eg, exercises that help recovery, suitable foods to eat, or things that the user needs to avoid). If the user's condition constantly worsens, the system will automatically suggest that the user should consult an actual doctor. In the event of an emergency, the user can press the emergency button, and the system will match the user with the fastest medical assistance available. If the user's physical condition becomes stable for a certain period of time, the system will inform the user that they have recovered and can go outside.
- *Personal recommendations*: Monitor users' mood based on the recording of their health condition, voice diaries, and interactions with the app. Post examples of users with mild symptoms to show a high possibility of full recovery and make them feel positive. In the meantime, inform users with real cases of COVID-19 about what they might experience in the days ahead and how to prepare themselves for it. For example, their health condition may worsen or fluctuate over the next few weeks. Based on the keywords retrieved by the users and the content viewed, combined with the health condition record, post positive information when signals of anxiety appear. When the user's condition has just improved, remind them that they still need to be careful and take it seriously.

Community Support

In this concept, users can socialize with those who have similar experiences online to get more social support.

- *Together with families*: With the consent of the user, share the users' health and emotional condition wheir families in case of emergency to enhance their feeling of connection.
- *Peer and community wisdom*: Increased socialization while helping each other by encouraging users to post their experiences and feelings, answer questions, and participate in a discussion group. In addition, a specific "meme module" can be provided to give users a chance to reduce their stress by sharing jokes, expressing their plight, and fostering empathy. Moreover, inspire users to try new hobbies that are shared by others on the hobby discussion board to reduce their boredom. Additionally, health care experts can participate to validate the posts. Rank the videos separately based on feedback from experts and other patients.

Regarding the development of an eHealth application using AI technology and its adaptation to the continuously changing situation of mild COVID-19, we recommend that application developers should add new concepts based on an existing eHealth application. By making a preliminary prototype and validating it with a small group of patients with mild COVID-19, developers should quickly iterate to meet missing needs that have not been considered before. It is necessary to be flexible based on how the COVID-19 situation develops and as regulations are updated.

Limitations and Implications for Further Research

Although the patient journey mapping is grounded on rigorous and systematic analysis of the qualitative data on the experiences of multiple persons with mild COVID-19 symptoms, this study has several limitations. In this design research, we used videos shared by people on YouTube as the main data source. The advantage of this self-shared data is that these patients have not been influenced by the researcher in advance, and the data is guaranteed to represent the patient perspective, which to a certain extent led the amount of information to exceed the researchers' expectations. The disadvantage is that unilateral dialogue without questions from researchers also meant that much of the data was irrelevant to the research question, which required the researchers to spend more time on sampling, extracting, and managing the risk that the personal video stories would not provide in-depth answers to some of the subquestions on the patients' experiences. In this regard, future research that includes additional face-to-face verification procedures is recommended to further enhance the robustness and reliability of the results. Concerning the sample strategy, the limitation of the current sample is that the participants are all YouTubers from the Americas and Europe, who tend to actively share and seek social support more easily than others. Further limitations relate to the racial and socioeconomic status disparities in online narratives that have been documented; in particular, stories by minorities are underrepresented on the internet, including on YouTube [47]. For future research and eHealth service design, more types of personal experience need to be considered. In addition, compared with other videos that are not shared on the

social media platform, YouTube videos have comments from viewers under each video. From those comments, we could easily collect different viewers' opinions on the video content, and we gained insights by analyzing those data. As most comments are composed of short sentences, future research could include AI technologies such as natural language processing and machine learning to efficiently analyze a larger number of comments.

Conclusions

The design of the patient journey map and the underlying insights into the home isolation experience serve to uncover new knowledge and enhance the professional understanding of persons with mild COVID-19 symptoms. The journey mapping synthesized urgent needs for eHealth service touchpoints, for instance, that patients require reliable public health information, personalized health monitoring guidance, and social support. To overcome the inadequate service provision challenges that became apparent in mapping the journey, initial service concepts were proposed for new AI eHealth services to improve the experience of patients with COVID-19 by providing effective health care guidance.

Acknowledgments

We would like to acknowledge all contributions of the participants in this study, which resulted in the patient journey map. Furthermore, we thank the reviewers for their constructive feedback that improved the quality of this paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Personal video story coverage and experienced symptoms during home isolation.
[PNG File , 2719 KB - medinform_v9i4e23238_app1.png]

Multimedia Appendix 2
Patient journey map of persons with mild COVID-19 during home isolation.
[PNG File , 782 KB - medinform_v9i4e23238_app2.png]

Multimedia Appendix 3
Video purpose and comments coding trees.
[DOCX File , 703 KB - medinform_v9i4e23238_app3.docx]

Multimedia Appendix 4
Visual summary of design research.
[PNG File , 1646 KB - medinform_v9i4e23238_app4.png]

References

1. Coronavirus disease (COVID-19) pandemic. World Health Organization. 2020 Jun 29. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 [accessed 2020-06-29]
2. WHO Coronavirus Disease (COVID-19) Dashboard. World Health Organization.: World Health Organisation; 2020 Jun 29. URL: https://covid19.who.int/ [accessed 2020-06-29]
3. Coronavirus disease (COVID-19): situation report– 160. World Health Organization. 2020 Jun 28. URL: https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200628-covid-19-sitrep-160.pdf?sfvrsn=2fe1c658_2 [accessed 2020-06-28]
4. Ebrahim SH, Ahmed QA, Gozzer E, Schlagenhauf P, Memish ZA. Covid-19 and community mitigation strategies in a pandemic. BMJ 2020 Mar 17;368:m1066. [doi: 10.1136/bmj.m1066] [Medline: 32184233]
5. Emanuel EJ, Persad G, Upshur R, Thome B, Parker M, Glickman A, et al. Fair allocation of scarce medical resources in the time of Covid-19. N Engl J Med 2020 May 21;382(21):2049-2055. [doi: 10.1056/NEJMsb2005114] [Medline: 32202722]
6. From New York to Tehran, "cabin hospitals" have been set up in many places around the world. Beijing Daily News. 2020 Mar 24. URL: http://ie.bjd.com.cn/5b165687a010550e5ddc0e6a/contentApp/5b16573ae4b02a9fe2d558f9/AP5e79b89ce4b0f99f4df7ce54?isshare=1 [accessed 2020-03-24]
7. Wong K. Wuhan puts together makeshift 'square cabin' hospitals in one night to treat mild coronavirus cases. Mothership. 2020 Feb 06. URL: https://mothership.sg/2020/02/wuhan-square-cabin-hospital/ [accessed 2020-02-06]
8. Pinedo E, Landauro I, Faus J. Spain to treat thousands of coronavirus patients in conference hall as toll tops 1,000. Reuters. 2020 Mar 20. URL: https://www.reuters.com/article/us-health-coronavirus-spain/

spain-to-treat-thousands-of-coronavirus-patients-in-conference-hall-as-toll-tops-1000-idUSKBN2172KO [accessed 2020-03-20]
9. Home care for patients with suspected novel coronavirus (nCoV) infection presenting with mild symptoms and management of contacts: interim guidance, 20 January 2020. World Health Organization. 2020 Jan 20. URL: https://apps.who.int/iris/handle/10665/330671 [accessed 2020-01-20]
10. Critical preparedness, readiness and response actions for COVID-19: interim guidance, 22 March 2020. World Health Organization. 2020 Mar 22. URL: https://apps.who.int/iris/handle/10665/331511?search-result=true&query=Critical+preparedness%2C+readiness+and+response+actions+for+COVID-19&scope=&rpp=10&sort_by=score&order=desc [accessed 2020-03-22]
11. Gandhi RT, Lynch JB, Del Rio C. Mild or moderate Covid-19. N Engl J Med 2020 Oct 29;383(18):1757-1766. [doi: 10.1056/NEJMcp2009249] [Medline: 32329974]
12. Folegatti PM, Bittaye M, Flaxman A, Lopez FR, Bellamy D, Kupke A, et al. Safety and immunogenicity of a candidate Middle East respiratory syndrome coronavirus viral-vectored vaccine: a dose-escalation, open-label, non-randomised, uncontrolled, phase 1 trial. Lancet Infect Dis 2020 Jul;20(7):816-826 [FREE Full text] [doi: 10.1016/S1473-3099(20)30160-2] [Medline: 32325038]
13. Tanne JH, Hayasaki E, Zastrow M, Pulla P, Smith P, Rada AG. Covid-19: how doctors and healthcare systems are tackling coronavirus worldwide. BMJ 2020 Mar 18;368:m1090. [doi: 10.1136/bmj.m1090] [Medline: 32188598]
14. Novel Coronavirus guidance: what to do when in doubt. Ministero della Salute. 2020 Mar 27. URL: http://www.salute.gov.it/portale/nuovocoronavirus/dettaglioOpuscoliNuovoCoronavirus.jsp?lingua=english&id=452 [accessed 2020-03-29]
15. Coronavirus disease (COVID-19): symptoms and treatment. Government of Canada. 2020 Mar 29. URL: https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/symptoms.html?topic=tilelink [accessed 2020-03-29]
16. Ik ben verkouden. Wat nu? Rijksinstituut voor Volksgezondheid en Milieu Ministerie van Volksgezondheid. 2020 Mar 25. URL: https://lci.rivm.nl/verkouden [accessed 2020-03-29]
17. Should you get tested. Centers for Disease Control and Prevention. 2020 Mar 21. URL: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html [accessed 2020-03-29]
18. Ohannessian R, Duong TA, Odone A. Global telemedicine implementation and integration within health systems to fight the COVID-19 pandemic: a call to action. JMIR Public Health Surveill 2020 Apr 02;6(2):e18810 [FREE Full text] [doi: 10.2196/18810] [Medline: 32238336]
19. Fagherazzi G, Goetzinger C, Rashid MA, Aguayo GA, Huiart L. Digital health strategies to fight COVID-19 worldwide: challenges, recommendations, and a call for papers. J Med Internet Res 2020 Jun 16;22(6):e19284 [FREE Full text] [doi: 10.2196/19284] [Medline: 32501804]
20. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. JMIR Ment Health 2020 Mar 26;7(3):e18848 [FREE Full text] [doi: 10.2196/18848] [Medline: 32213476]
21. Xu C, Zhang X, Wang Y. Mapping of health literacy and social panic via web search data during the COVID-19 public health emergency: infodemiological study. J Med Internet Res 2020 Jul 02;22(7):e18831 [FREE Full text] [doi: 10.2196/18831] [Medline: 32540844]
22. 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]
23. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial Intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr 2020;14(4):337-339 [FREE Full text] [doi: 10.1016/j.dsx.2020.04.012] [Medline: 32305024]
24. Kuziemsky C, Maeder AJ, John O, Gogia SB, Basu A, Meher S, et al. Role of artificial intelligence within the telehealth domain. Yearb Med Inform 2019 Aug;28(1):35-40 [FREE Full text] [doi: 10.1055/s-0039-1677897] [Medline: 31022750]
25. Keesara S, Jonas A, Schulman K. Covid-19 and health care's digital revolution. N Engl J Med 2020 Jun 04;382(23):e82. [doi: 10.1056/NEJMp2005835] [Medline: 32240581]
26. Tran BX, Nghiem S, Sahin O, Vu TM, Ha GH, Vu GT, et al. Modeling research topics for artificial intelligence applications in medicine: latent Dirichlet allocation application study. J Med Internet Res 2019 Nov 01;21(11):e15511 [FREE Full text] [doi: 10.2196/15511] [Medline: 31682577]
27. D'Alfonso S. AI in mental health. Curr Opin Psychol 2020 Dec;36:112-117. [doi: 10.1016/j.copsyc.2020.04.005] [Medline: 32604065]
28. Bhattad PB, Jain V. Artificial intelligence in modern medicine - the evolving necessity of the present and role in transforming the future of medical care. Cureus 2020 May 09;12(5):e8041 [FREE Full text] [doi: 10.7759/cureus.8041] [Medline: 32528777]
29. Kueper JK, Terry AL, Zwarenstein M, Lizotte DJ. Artificial intelligence and primary care research: a scoping review. Ann Fam Med 2020 May;18(3):250-258 [FREE Full text] [doi: 10.1370/afm.2518] [Medline: 32393561]
30. Bucci S, Schwannauer M, Berry N. The digital revolution and its impact on mental health care. Psychol Psychother 2019 Jun;92(2):277-297. [doi: 10.1111/papt.12222] [Medline: 30924316]

31. Bernal G, Colombo S, Al Ai Baky M, Casalegno F. Safety++ designing IoT and wearable systems for industrial safety through a user centered design approach. In: Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments. 2017 Presented at: PETRA '17; June 2017; Island of Rhodes, Greece p. 163-170. [doi: 10.1145/3056540.3056557]
32. Xu W. Toward human-centered AI: a perspective from human-computer interaction. Interactions 2019 Jun 26;26(4):42-46. [doi: 10.1145/3328485]
33. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: systematic review. J Med Internet Res 2020 Feb 20;22(2):e16866 [FREE Full text] [doi: 10.2196/16866] [Medline: 32130134]
34. Riedl MO. Human‐centered artificial intelligence and machine learning. Hum Behav Emerging Technologies 2019 Feb 07;1(1):33-36. [doi: 10.1002/hbe2.117]
35. Asan O, Bayrak AE, Choudhury A. Artificial intelligence and human trust in healthcare: focus on clinicians. J Med Internet Res 2020 Jun 19;22(6):e15154 [FREE Full text] [doi: 10.2196/15154] [Medline: 32558657]
36. Tran BX, Latkin CA, Sharafeldin N, Nguyen K, Vu GT, Tam WWS, et al. Characterizing artificial intelligence applications in cancer research: a latent Dirichlet allocation analysis. JMIR Med Inform 2019 Sep 15;7(4):e14401 [FREE Full text] [doi: 10.2196/14401] [Medline: 31573929]
37. De Silva D, Ranasinghe W, Bandaragoda T, Adikari A, Mills N, Iddamalgoda L, et al. Machine learning to support social media empowered patients in cancer care and cancer treatment decisions. PLoS One 2018;13(10):e0205855 [FREE Full text] [doi: 10.1371/journal.pone.0205855] [Medline: 30335805]
38. Ribera M, Lapedriza A. Can we do better explanations? A proposal of User-Centered Explainable AI. IUI Workshops. 2019. URL: http://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf [accessed 2020-04-01]
39. Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. 2019 Presented at: CHI '19; May 2019; Glasgow, Scotland, UK p. 1-15. [doi: 10.1145/3290605.3300831]
40. Braun V, Clarke V. Successful Qualitative Research: A Practical Guide for Beginners. Los Angeles, CA: Sage; 2013.
41. Van Manen M. Researching Lived Experience: Human Science for an Action Sensitive Pedagogy. London: Routledge; 2016.
42. Simonse LWL, Albayrak A, Starre S. Patient journey method for integrated service design. Design Health 2019 May 13;3(1):82-97. [doi: 10.1080/24735132.2019.1582741]
43. Micheli P, Wilner SJS, Bhatti SH, Mura M, Beverland MB. Doing design thinking: conceptual review, synthesis, and research agenda. J Prod Innov Manag 2018 Sep 08;36(2):124-148. [doi: 10.1111/jpim.12466]
44. McCarthy SO, O'Raghallaigh P, Woodworth S, Lim YL, Kenny LC, Adam F. An integrated patient journey mapping tool for embedding quality in healthcare service reform. J Decision Syst 2016 Jun 16;25(sup1):354-368. [doi: 10.1080/12460125.2016.1187394]
45. Trebble TM, Hansi N, Hydes T, Smith MA, Baker M. Process mapping the patient journey: an introduction. BMJ 2010 Aug 13;341:c4078. [doi: 10.1136/bmj.c4078] [Medline: 20709715]
46. Ben-Tovim DI, Dougherty ML, O'Connell TJ, McGrath KM. Patient journeys: the process of clinical redesign. Med J Aust 2008 Mar 17;188(S6):S14-S17. [doi: 10.5694/j.1326-5377.2008.tb01668.x] [Medline: 18341470]
47. Chou WS, Hunt Y, Folkers A, Augustson E. Cancer survivorship in the age of YouTube and social media: a narrative analysis. J Med Internet Res 2011 Jan 17;13(1):e7 [FREE Full text] [doi: 10.2196/jmir.1569] [Medline: 21247864]
48. Grajales FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. J Med Internet Res 2014 Feb 11;16(2):e13 [FREE Full text] [doi: 10.2196/jmir.2912] [Medline: 24518354]
49. Miles MB, Huberman MA, Saldaña J. Qualitative Data Analysis: A Methods Sourcebook. Newbury Park, CA: SAGE Publications Inc; 2013.
50. Karan A. To control the covid-19 outbreak, young, healthy patients should avoid the emergency department. BMJ 2020 Mar 17;368:m1040. [doi: 10.1136/bmj.m1040] [Medline: 32184232]
51. Polkinghorne DE. Phenomenological research methods. In: Valle RS, Halling S, editors. Existential-Phenomenological Perspectives in Psychology: Exploring the Breadth of Human Experience. Boston, MA: Springer; 1989:41-60.
52. Human Research Ethics. Delft University of Technology. URL: https://www.tudelft.nl/en/about-tu-delft/strategy/integrity-policy/human-research-ethics/ [accessed 2019-03-13]
53. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013 Sep;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
54. Ravitch SM, Carl MNC. Qualitative Research: Bridging the Conceptual, Theoretical, and Methodological. Newbury Park, CA: SAGE Publications Inc; 2015.
55. Saldaña J. The Coding Manual for Qualitative Researchers. London: SAGE; 2012.
56. Sanders EBN, Stappers PJ. Convivial Toolbox. Amsterdam, NL: BIS Publishers; 2013.

Abbreviations

AI: artificial intelligence
CEST: Central European Summer Time
WHO: World Health Organization

Edited by G Eysenbach; submitted 05.08.20; peer-reviewed by W Yaogang, O Ogundaini; comments to author 18.11.20; revised version received 18.12.20; accepted 10.01.21; published 12.04.21.

Please cite as:
He Q, Du F, Simonse LWL
A Patient Journey Map to Improve the Home Isolation Experience of Persons With Mild COVID-19: Design Research for Service Touchpoints of Artificial Intelligence in eHealth
JMIR Med Inform 2021;9(4):e23238
URL: https://medinform.jmir.org/2021/4/e23238
doi:10.2196/23238
PMID:33444156

Original Paper

Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development

Sakifa Aktar[1*], BSc; Md Martuza Ahamad[1*], MSc; Md Rashed-Al-Mahfuz[2], MSc; AKM Azad[3], PhD; Shahadat Uddin[4], PhD; AHM Kamal[5], PhD; Salem A Alyami[6], PhD; Ping-I Lin[7], PhD; Sheikh Mohammed Shariful Islam[8], PhD; Julian MW Quinn[9], PhD; Valsamma Eapen[7], PhD; Mohammad Ali Moni[7,9,10], PhD

[1]Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj, Bangladesh

[2]Department of Computer Science and Engineering, University of Rajshahi, Rajshahi, Bangladesh

[3]iThree Institute, Faculty of Science, University Technology of Sydney, Sydney, Australia

[4]Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Darlington, Sydney, Australia

[5]Department of Computer Science and Engineering, Jatiya Kabi Kazi Nazrul Islam University, Mymensingh, Bangladesh

[6]Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

[7]School of Psychiatry, Faculty of Medicine, University of New South Wales, Sydney, Australia

[8]Institute for Physical Activity and Nutrition, Faculty of Health, Deakin University, Victoria, Australia

[9]Healthy Ageing Theme, The Garvan Institute of Medical Research, Darlington, Australia

[10]WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, University of New South Wales, Sydney, Australia

[*]these authors contributed equally

Corresponding Author:
Mohammad Ali Moni, PhD
WHO Collaborating Centre on eHealth, UNSW Digital Health
School of Public Health and Community Medicine, Faculty of Medicine
University of New South Wales
Kensington
Sydney, NSW 2052
Australia
Phone: 61 414701759
Email: m.moni@unsw.edu.au

Abstract

Background: Accurate prediction of the disease severity of patients with COVID-19 would greatly improve care delivery and resource allocation and thereby reduce mortality risks, especially in less developed countries. Many patient-related factors, such as pre-existing comorbidities, affect disease severity and can be used to aid this prediction.

Objective: Because rapid automated profiling of peripheral blood samples is widely available, we aimed to investigate how data from the peripheral blood of patients with COVID-19 can be used to predict clinical outcomes.

Methods: We investigated clinical data sets of patients with COVID-19 with known outcomes by combining statistical comparison and correlation methods with machine learning algorithms; the latter included decision tree, random forest, variants of gradient boosting machine, support vector machine, k-nearest neighbor, and deep learning methods.

Results: Our work revealed that several clinical parameters that are measurable in blood samples are factors that can discriminate between healthy people and COVID-19–positive patients, and we showed the value of these parameters in predicting later severity of COVID-19 symptoms. We developed a number of analytical methods that showed accuracy and precision scores >90% for disease severity prediction.

Conclusions: We developed methodologies to analyze routine patient clinical data that enable more accurate prediction of COVID-19 patient outcomes. With this approach, data from standard hospital laboratory analyses of patient blood could be used to identify patients with COVID-19 who are at high risk of mortality, thus enabling optimization of hospital facilities for COVID-19 treatment.

(JMIR Med Inform 2021;9(4):e25884) doi:10.2196/25884

KEYWORDS

COVID-19; blood samples; machine learning; statistical analysis; prediction; severity; mortality; morbidity; risk; blood; testing; outcome; data set

Introduction

SARS-CoV-2 has caused the current pandemic of COVID-19, a disease that first emerged as an outbreak in December 2019 in the Chinese province of Hubei [1]. The management of patients with COVID-19 remains problematic and controversial, although this is to be expected in such a recently emerged disease. The first symptoms of COVID-19 resemble those of many other infections and inflammatory conditions that affect the respiratory system; they include fever, sneezing and rhinitis, persistent cough, and fatigue with body ache [2]. However, an infected patient can rapidly develop additional and more severe symptoms that can be life-threatening and require intensive care intervention; these include pneumonia, severe shortness of breath, diarrhea, dispersed thrombosis, and vascular inflammation [3,4]. An additional issue in caring for patients with COVID-19 is the presence of comorbidities that interact with COVID-19, particularly pulmonary and vascular conditions, which can greatly worsen the patient's prognosis [5]. This is an important consideration given the current lack of effective therapy for COVID-19. However, there have been notable advances in treating patients with advanced disease; therefore, the ability to predict that a patient will have poor outcomes, indicating a need for more aggressive treatment, has the potential to save lives and enable more effective allocation of resources.

Intensive care units (ICUs) are key to increasing the survival of patients with severe COVID-19; they provide oxygen, 24-hour monitoring and care, and assisted ventilation when needed. Therefore, ICU beds are a precious resource in locations where COVID-19 case numbers are high [6-8]. Allocating hospital wards or ICU beds for infected patients thus requires rapid decision-making processes, both to use resources efficiently and reduce patient suffering and mortality. In many parts of the world, stressed care systems face significant difficulty in deciding on ICU bed allocation; therefore, a smart, automated system could be useful to improve care and resource allocation. The World Health Organization has recommended that all suspected patients with COVID-19 be tested by reverse transcription–polymerase chain reaction (RT-PCR)–based diagnosis methods that directly detect viral RNA [9]. Testing by approaches other than RT-PCR does not yet show acceptable accuracy. However, RT-PCR tests can take many hours or days to finalize the test outcomes, by which time the health condition and infectious status of confirmed patients may deteriorate. Rather than seeking a new single rapid test that improves on RT-PCR, an alternative approach could be to use results from many different profiling tests that are already available and can be performed quickly using existing equipment [10,11]. The best way to use the resulting multidimensional data is currently controversial.

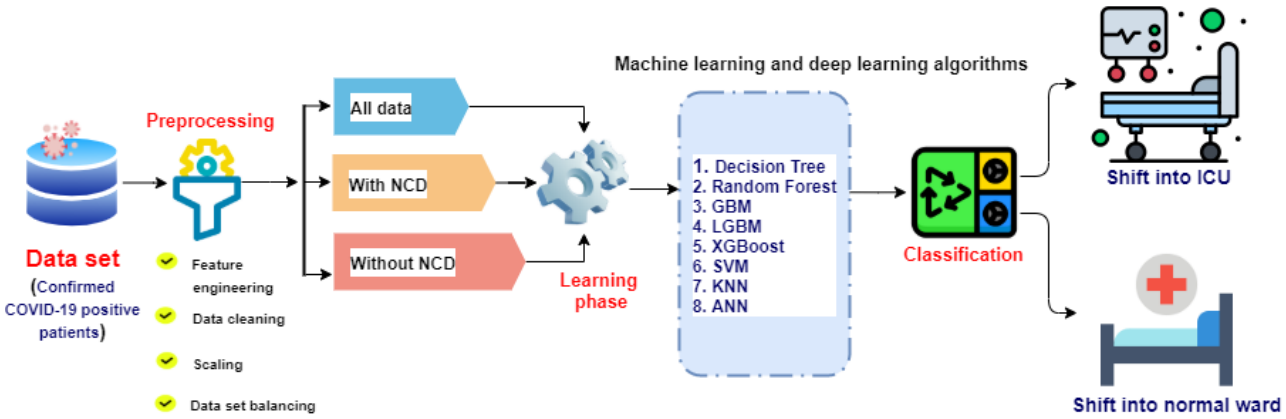
Rapid blood and serology testing of clinical samples by current equipment enables monitoring of many peripheral blood parameters of interest, some of which indicate changes in organ functions and are used to diagnose a range of conditions and diseases [7,12]. This raises the possibility that such profiling of blood samples could provide predictive information about the disease trajectory and risk of comorbidities for patients with COVID-19. Some data is already used in physician deliberations; however, the many available test parameters suggest that an agnostic statistical or machine learning (ML) approach would improve the quality of those decisions. Therefore, we undertook a comprehensive assessment that examined the utility of a range of statistical and ML approaches. Indeed, we identified algorithms that showed significantly improved outcome estimates. Therefore, this work has the potential to optimize decision processes regarding patient care by clinicians who are under significant time and resource pressure during the current COVID-19 pandemic.

Methods

Data Sets and Analyses

We used two different data sets in this study; the first included data from 89 patients, and the second included data from 1945 patients with confirmed positive COVID-19 tests identified by RT-PCR. For the first data set [13], we use statistical methods such as the Student *t* test, chi-square test, and Pearson correlation to identify the most significant and associative blood parameters that can strongly distinguish between patients with COVID-19 and healthy people. Moreover, to compare the blood parameter values of patients with COVID-19 with those of healthy patients, we considered the standard value ranges as reference values for each parameter. For the second data set [14], in addition to statistical methods, we used several ML models to further identify blood parameters that can discriminate between COVID-19–positive patients who are at risk of serious illness and those who are not. Figure 1 depicts a schematic of the ML analysis workflow of our approach.

Figure 1. Proposed methodology and workflow of the machine learning analysis in this study. ANN: artificial neural network; GBM: gradient boosting machine; ICU: intensive care unit; LGBM: light gradient boosting machine; NCD: noncommunicable disease; SVM: support vector machine; KNN: k-nearest neighbor; XGBoost: extreme gradient boosting.



We formulated the task of identifying patients with severe COVID-19 to enable selection of the appropriate hospital ward for their care as a classification problem by training ML models with features of clinical data collected from blood samples of patients with COVID-19. Raw data of interest collected from the data sets underwent a data-wrangling pipeline, including denoising, missing value imputation, transformation, normalization, and partition. Next, several statistical comparisons and correlation methods were adopted for feature engineering, including the Student *t* test, chi-square test, and Pearson correlation. After this, each data set was further split into three categories based on the criteria of existing noncommunicable disease (NCD): with NCD, without NCD, and all data. In our study, "NCD" refers to patients with pre-existing noncommunicable diseases or conditions. Finally, a range of state-of-the-art ML methods were trained and evaluated. The algorithms used included decision tree (DT), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), support vector machine (SVM), light gradient boosting machine (LGBM), k-nearest neighbor (KNN), and artificial neural network (ANN)–based deep learning sequential models. Each of these steps is discussed in the following subsections.

Data Collection

We obtained two different data sets of patients with COVID-19. The first data set was produced by Zenodo [13], and it contains demographic information and blood sample information from 89 COVID-19–positive patients. In this data set, 31 patients were alive at the point of data collection, while 58 patients had died. The second, larger data set was obtained from the Kaggle web-based resource [14], which contains grouped information regarding previous diseases, blood sample results, and vital sign data of 1945 COVID-19–positive patients. The primary sources of the data in this set are Brazilian hospitals, including Sirio Libanes, São Paulo, and Brasilia. The parameters of the data set included patient age percentile, gender, and demographic information. Some patients had pre-existing NCDs, including hypertension and immunocompromised status. The blood parameters examined included lactate, respiratory rate, diastolic blood pressure, hemoglobin, hematocrit, venous base excess, leukocytes, neutrophils, albumin, arterial base excess, urea, platelets, potassium, systolic blood pressure, venous PO_2, arterial O_2 saturation, partial thromboplastin time, temperature, gamma-glutamyl transferase, venous O_2 saturation, creatinine, international normalized ratio (INR), venous PCO_2, venous pH, arterial bicarbonate, labels of free fatty acids, venous bicarbonate, calcium, lymphocytes, alanine aminotransferase, aspartate aminotransferase, arterial PCO_2, dimerized plasmin fragment D (D-dimer), oxygen saturation, bilirubin, arterial PO_2, arterial pH, heart rate, blast, and glucose. During the feature-engineering phase in our study, all these blood parameters were considered as features.

Data Processing

For the Zenodo data set [13], which consists of 89 COVID-19–positive patients, we first removed any unwanted parameters (eg, ethnicity, BMI, drinking or smoking habits). We then eliminated all the missing values, resulting in a data set of 70 patients. In the Sirio Libanes data set [14] from Kaggle, there were 1945 individual patients with 54 types of tests. The primary data set contained a large number of missing values. This data set was prepared from information received from local hospitals and some of this information was not well prepared, which is a significant reason why most of the data have missing entries. The rationale behind the removal of entries with missing parameter values is that when we conducted a pilot study with the imputation of missing values with mean, median, or regression values, poor predictive performance was observed. In the raw data set, the dimensions were 1925 × 205, and almost 57% of the data units (cell values) were missing; after eliminating unwanted attributes, the amount of missing data increased above 70%. If we considered all the data and imputed the missing values, most of the values would be inferred, and the analysis results would be unreliable. Therefore, we eliminated entries that contained at least one missing value. This elimination resulted in 545 sets of patient data entries in the second data set that contained no missing values. Among the patients in this data set, 264 had sufficiently severe symptoms to be admitted to the ICU. Both data sets underwent a denoising step, in which we removed unwanted strings. Standard scaling techniques were performed, such as feature scaling, in which the variance values of the data are scaled between 0 and 1; this is calculated by subtracting the mean value

of a feature from the original value and then dividing by the standard deviation. After preprocessing, we considered data from 545 patients for the analysis. For a precise study, we then divided this data set according to whether a patient had a coexisting NCD (NCD) or not (no NCD). We found 264 patients with NCDs and 281 patients without NCDs; in the NCD and no NCD groups, 156 and 108 patients were respectively classed as displaying severe conditions. After this data preparation and preprocessing, we considered all these data for the statistical analysis. Due to the possibility of data leakage in ML analysis if we separated the test set and train sets after preprocessing, we first separated a randomly selected 80% of the grouped patient data for model training and used the rest for model validation testing, then performed the preprocessing steps.

Statistical Methods to Identify the Most Significant and Associative Blood Parameters

In the statistical analysis, we used chi-square tests for categorical variables, Student *t* tests for continuous variables, and Pearson correlations among various blood sample counts. The null hypothesis was that the data from the patients with COVID-19 and the healthy population were independent. Significant blood parameters were chosen based on a *P* value <.05, while in some cases, the selection criteria were a false discovery rate–adjusted *P* value <.05 and an absolute value log 2 fold change (LFC) <1. To understand the changes (positive or negative) of the parameters and the number of changes, we have calculated the LFC. LFC=1 indicates a fold change of value 2. Furthermore, hierarchical clustering was conducted on the Pearson correlation coefficients for grouping significant parameters [15-17].

ML Models to Classify COVID-19 Disease Severity

To identify a set of important blood samples as a feature selection step, we employed a set of ML algorithms using COVID-19 data sets that included data from severely and nonseverely affected patients. We chose ML algorithms that are known to perform classification tasks with superior performance and fast execution [18,19]. For this purpose, we considered a basic ensemble learning approach based on max-voting, averaging, and weighted averaging for some classifiers, as well as advanced ensemble learning algorithms that function by stacking, blending, bagging, and boosting. Ensemble learning algorithms are combinations of one or more basic algorithms that are high-performing, efficient, effective, and easy to debug [20,21].

We next address the parameters of the ML algorithms that were considered when they were run. In the DT algorithm, we used a random state of 42, a criterion of Gini, and a minimum sample split of 2. Similarly, in the RF algorithm, the minimum sample split was 2 and the number of estimators was 100. Degree and kernel cache size are parameters of the SVM algorithm; the algorithm sets a polynomial kernel with a degree of 3, and we set the kernel cache size at 200 MB for fast execution. In the GBM algorithm, the learning rate was 0.1, the criterion was friedman_mse, and the number of estimators was 100. The learning rate in the LGBM algorithm was 0.05, the feature fraction was 0.9, the bagging fraction was 0.8, and the bagging frequency was 5. In the XGB algorithm, we used a tree-based booster with a maximum depth of 6, a learning rate of 0.1, and 1000 estimators. For the KNN algorithms, we used Minkowski matrices; the weights were uniform, and the number of neighbors was 3 (k=3).

We also experimented with a sequential deep learning model, namely, a feed-forward 1D ANN. This model consists of an input layer, three hidden layers, and an output layer [22]. Each layer contains a collection of parallel processing nodes, called neurons, that take input from the nodes of the previous layer. All the hidden layers are activated by rectified linear units, and the output layer is activated by a softmax function, providing the class probability of the input sample. The network was trained in 1000 epochs using the stochastic gradient descent optimization algorithm with categorical cross-entropy loss as a convergence indicator and a learning rate of 0.0001.

Shapley Additive Explanation Value Calculations

To measure the feature importance, we calculated the Shapley Additive Explanation (SHAP) values from all the models to estimate the degree of contribution of each of the features in the samples of the training data set to the overall decision-making of the model [23]. SHAP uses game theory rules to determine the contributions of particular features to the decision-making of the model. We used the TreeExplainer [24] for tree-based models and the KernelExplainer [23] for kernel-based models to calculate the feature importance. After finding the SHAP values for all the models, we normalized the values in a fixed range and considered the average values.

Evaluation Matrices for the ML Models

We evaluated the performance of our models using precision, recall, F1 score, the area under the receiver operator characteristic curve (AUC-ROC), and the log loss function. The precision depicts the proportion of true positive instances among all the predicted positive instances [25]; in contrast, the recall shows the proportion of the actual true instances that are predicted positively by the models [25]. The F1 score is the harmonic mean of precision and recall [25]; we calculated the F1 scores to achieve better evaluation between precision and recall. The AUC of a classifier is equivalent to the likelihood that the classifier will rank a randomly selected positive value higher than a randomly selected negative value [26]. Log loss is also essentially used as a metric for classification; it is calculated by the probability of actual and predicted classes [27]. Log loss is among the most useful evaluation metrics. The function can be described as below:

where M depicts the number of classes, T_i indicates the actual class, and $p(T_i)$ indicates the probability of that class.

Results

Analysis Approaches

In this study, we adopted two scenarios for analyzing research data. In the first scenario, we applied the Student *t* test and Pearson correlation to the blood cell parameters of COVID-19–positive patients and the normal ranges of the blood cell parameters. We found that both statistical approaches

yielded predictive capability of immature granulocytes (absolute), hemoglobin A_{1c}, fibrinogen, and lipase as significant for COVID-19–positive patients. In the second scenario, we accounted only for COVID-19–positive patients in the severity calculation. We also applied two different analysis approaches. The first one was the Student *t* test, and the second was a set of ML methods. Using both of these approaches, we found that respiratory rate, lactate, blood pressure (systolic and diastolic), hemoglobin, hematocrit, venous and arterial base excess, neutrophils, albumin, urea, platelet count, and potassium were good indicators of the patients' disease severity and represented a small set of predictors of COVID-19 severity measurements.

Patient Demographics

A comparison of the demographic information for the data from the patients with severe and nonsevere symptoms is shown in Table 1. This distribution table is included here to show the distribution of patients in the data set clearly. Of the 545 patients, 198 (36.3%) were female, 257 (47.2%) were above 65 years of age, and 264 (48.4%) were admitted to the ICU. Among the group that included only patients with no NCDs (n=281), 107 (38.1%) were female, and 108 (38.4%) were admitted to the ICU. Moreover, in the group of patients who had one or more NCDs (n=264), 167 (63.3%) were over 65 years of age, and 156 (59.1%) were admitted to the ICU. The age percentile is shown in Figure 2.

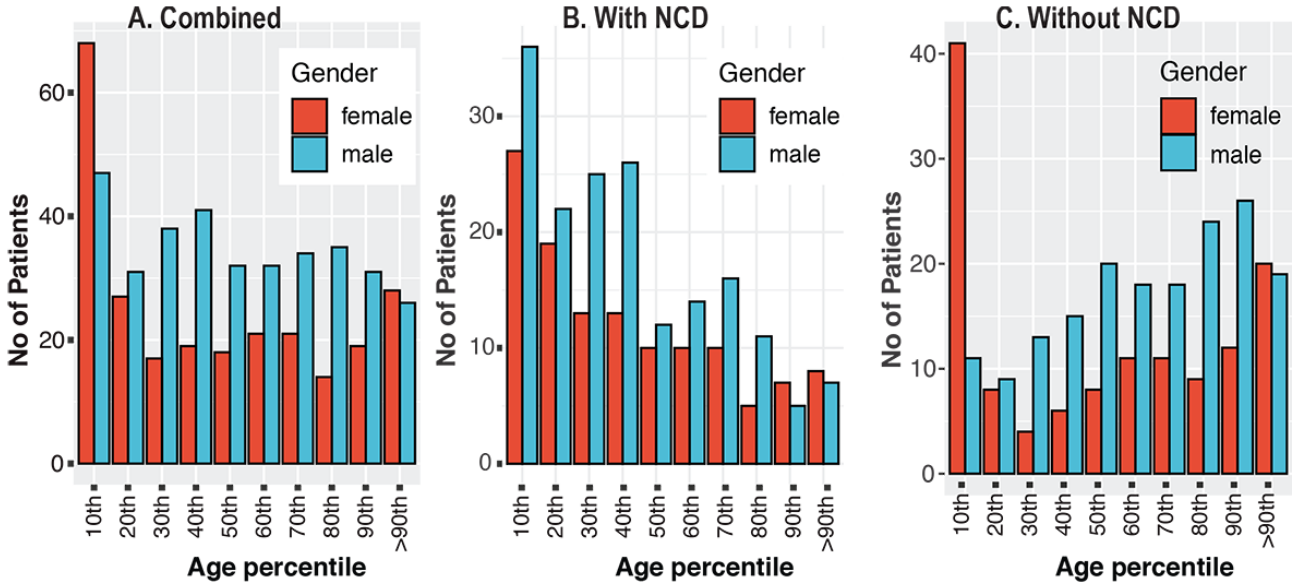
Table 1. Demographic information for the patients with COVID-19 in each patient group.

Characteristic	Values, n (%)		
	All patients (N=545)	Patients without NCDs[a] (n=281)	Patients with NCDs (n=264)
Age >65 years	257 (47.2)	90 (32.0)	167 (63.3)
Age percentile			
10th	115 (21.1)	63 (22.4)	52 (19.7)
20th	58 (10.6)	41 (14.6)	17 (6.4)
30th	55 (10.1)	38 (13.5)	17 (6.4)
40th	60 (11.0)	39 (13.9)	21 (8.0)
50th	50 (9.2)	22 (7.8)	28 (10.6)
60th	53 (9.7)	24 (8.5)	29 (11.0)
70th	55 (10.1)	26 (9.3)	29 (11.0)
80th	49 (9.0)	16 (5.7)	33 (12.5)
90th	50 (9.2)	12 (4.3)	38 (14.4)
>90th	54 (9.9)	15 (5.3)	39 (14.8)
Female gender	198 (36.3)	107 (38.1)	91 (34.5)
Admitted to ICU[b]	264 (48.4)	108 (38.4)	156 (59.1)

[a]NCDs: noncommunicable diseases.

[b]ICU: intensive care unit.

Figure 2. Age percentiles of patients with COVID-19 for (A) both patient groups, (B) patients with NCDs, and (C) patients without NCDs. NCD: noncommunicable disease.



Identification of Significant Routine Blood Parameters for SARS-CoV-2 Infection

Our first data set contained 89 blood parameters for confirmed COVID-19–positive patients. Assuming each blood parameter value was normally distributed in the healthy population, we performed Student *t* tests on the tested blood parameters to compare the expected range values (shown in Figure 3) with patients with COVID-19 from the first data set. The combination of Student *t* test and LFC analyses indicated that the 8 most significant candidate predictive parameters for COVID-19 severity status were lipase, C-reactive protein, procalcitonin level, erythrocyte sedimentation rate, brain natriuretic peptide, ferritin, D-dimer, and creatine kinase level, all of which showed *P* values <.001 and absolute LFCs >1.

We applied the Student *t* test to the second data set to attempt to discriminate symptoms of severe and nonsevere COVID-19–positive patients by identifying patient characteristics that are associated with the target variable of disease severity; the analysis results are shown in Figure 4. The most significant blood parameters according to the *t* test results were lactate, respiratory rate, diastolic blood pressure, hemoglobin, hematocrit, venous base excess, leukocytes, neutrophils, albumin, arterial base excess, urea, platelet count, potassium, and systolic blood pressure.

Figure 3. Parameter measurements for various blood parameters and significant differences (using *t* tests) between patients with and without COVID-19. Adj.p-value: adjusted *P* value; D-dimer: dimerized plasmin fragment D.

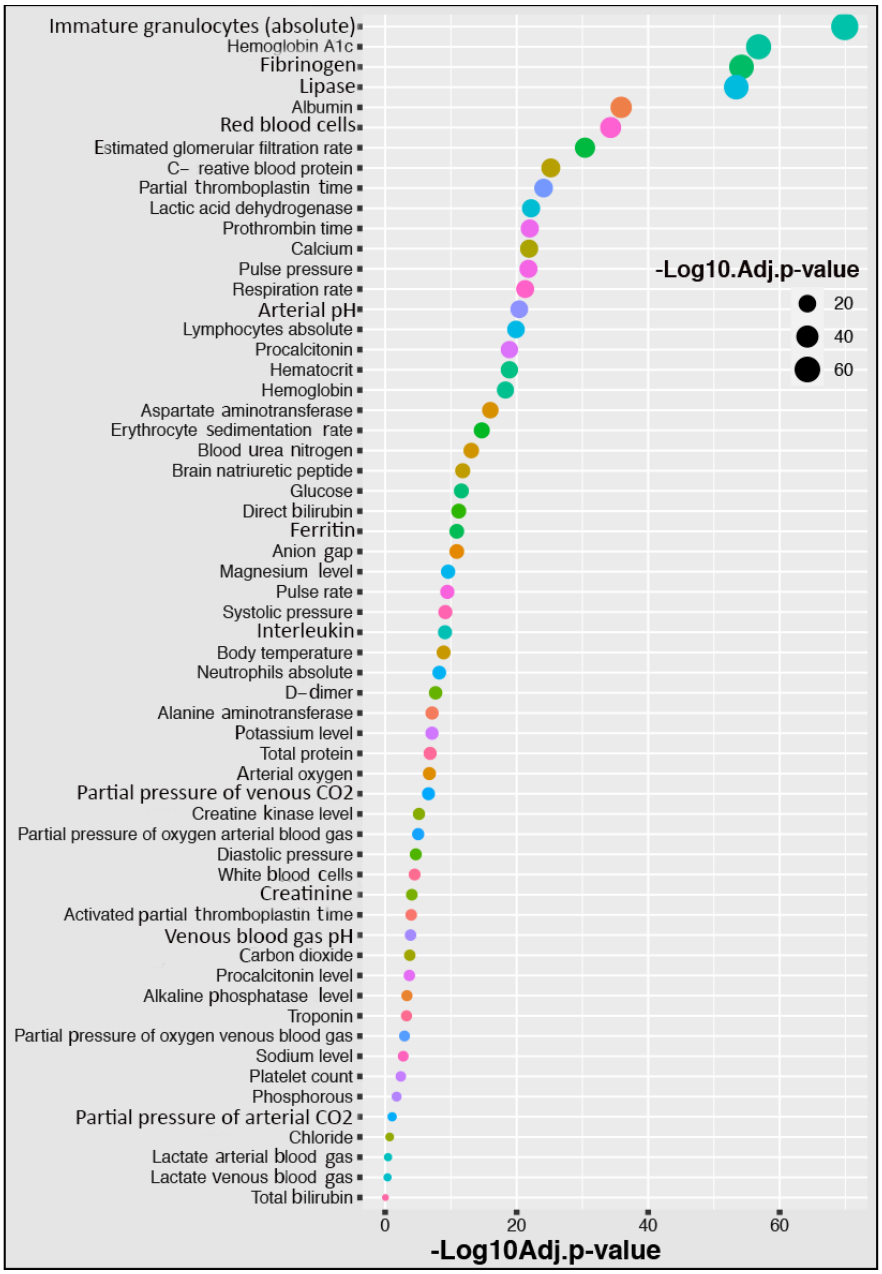
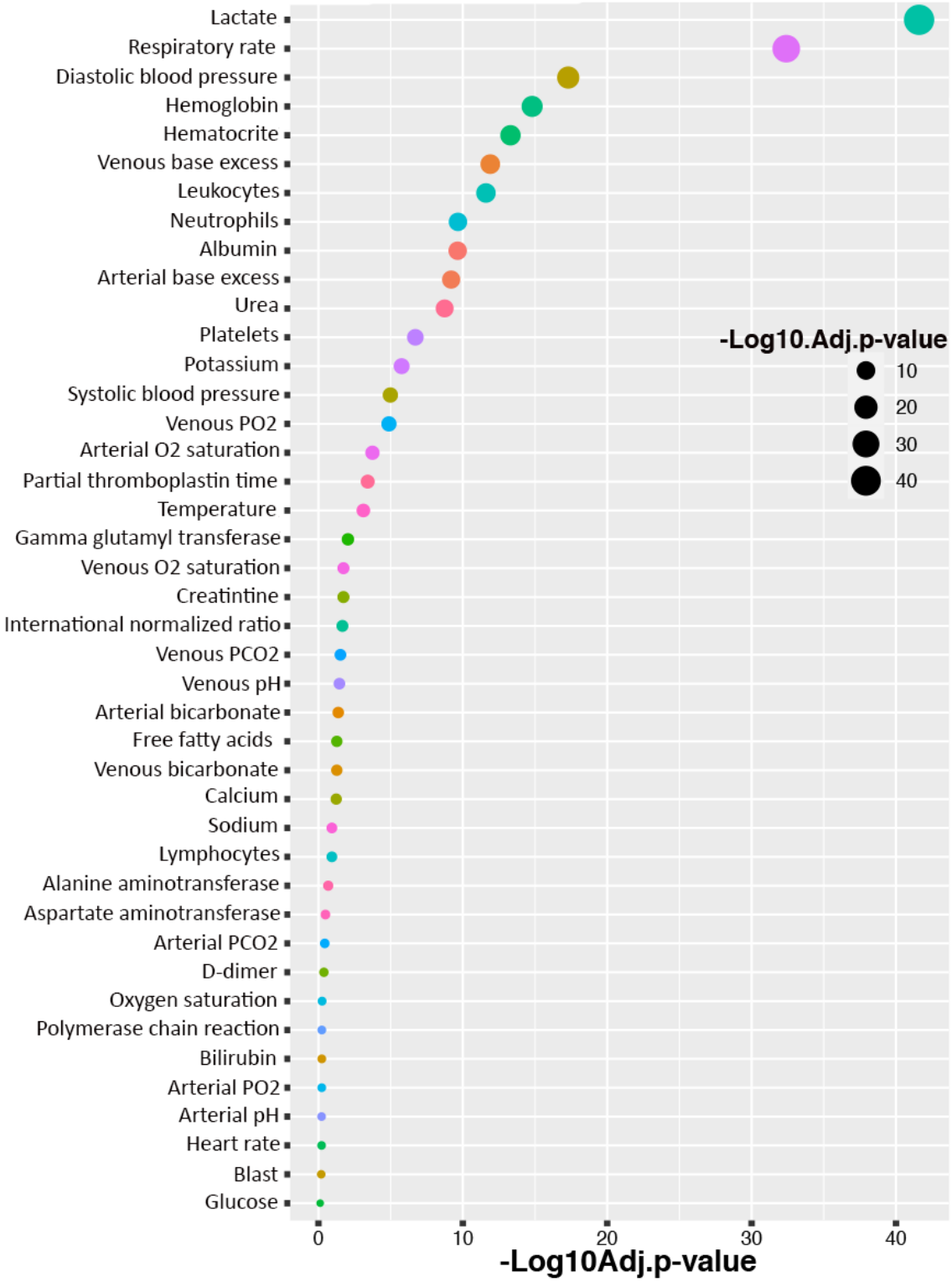


Figure 4. Association of blood parameters with the severity of COVID-19 disease. Associations and significant differences (using *t* tests) between the patients with severe COVID-19 and nonsevere COVID-19. Adj.p-value: adjusted *P* value; D-dimer: dimerized plasmin fragment D; FFA: free fatty acids; GGT: gamma-glutamyl transferase; INR: international normalized ratio.



Clustering and Coexpression Analysis

We also performed Pearson correlation tests for the different routine blood parameters. The Pearson correlation results are shown in Figure 5. The purpose of the hierarchical clustering was to observe which blood samples share similar properties in terms of their values among all the patients. We found that some blood features formed clusters, which indicates that they share similar properties among patients. We found that there were indeed some hierarchical clusters in the tests that showed equal significance for all the patients. From the total of 59 blood samples, we found 4 different concordant clusters that were strongly correlated with each other. The first cluster comprised pulse pressure and systolic blood pressure. The second cluster comprised hemoglobin, hematocrit, and red blood cells. The

third cluster comprised C-reactive protein, erythrocyte sedimentation rate, diastolic blood pressure, and respiratory rate. Procalcitonin levels, ferritin, and creatine kinase levels composed the fourth cluster.

Figure 5. Correlation heat map among the various blood parameters examined using the data set of 89 patients. D-dimer: dimerized plasmin fragment D.

Prediction of Severe COVID-19 for Critical Treatment Using ML Models

In this section, we first describe the performance of the various ML models employed and their applications. We then present the most important reduced set of blood and physical sign parameters that can precisely discriminate patients with severe COVID-19 from those with nonsevere disease. The reduced collection of blood parameters is also significant for outcomes of patients with severe COVID-19.

For the ML analysis of the second data set, we applied the respective methods and models; their performances and the evaluation matrices are shown in Table 2. In the data group of all patients with and without NCDs, we found that the RF and GBM methods gave the highest testing accuracy score of 89%, and the other methods and models demonstrated >80% testing accuracy. The highest AUC was obtained for RF and GBM (89%), and other methods and models achieved suitable AUC values >80%. The highest precision value of 91% was observed for XGB and GBM. The highest recall values obtained were 93% for KNN and 90% for RF and LGBM; the other methods showed scores above 80%. The best F1 score was 90% for RF, and the other models showed F1 scores >80%. RF and GBM had the lowest log loss value of 3.8%, and the other methods and models also showed particularly low values (ie, <7%). In this patient group, we saw that all of our applied models achieved good performance in every evaluation matrix with accuracy scores >80%; therefore, in practice, any of the models can be employed.

Table 2. Accuracy and evaluation matrices for each data group.

Data set and matrices	RF[a]	LGBM[b]	SVM[c]	DT[d]	XGB[e]	GBM[f]	KNN[g]	ANN[f]
Combined								
Accuracy	0.89	0.88	0.84	0.82	0.88	0.89	0.84	0.83
AUC[g]	0.89	0.88	0.84	0.82	0.88	0.89	0.84	0.82
Precision	0.9	0.88	0.84	0.83	0.91	0.91	0.81	0.92
Recall	0.9	0.9	0.88	0.83	0.86	0.88	0.93	0.69
F1 score	0.9	0.89	0.86	0.83	0.88	0.89	0.86	0.79
Log loss	3.8	4.12	5.39	6.34	4.12	3.8	5.39	6.02
With NCDs[h]								
Accuracy	0.91	0.93	0.84	0.84	0.87	0.89	0.77	0.74
AUC	0.91	0.92	0.83	0.84	0.87	0.89	0.79	0.71
Precision	0.89	0.89	0.83	0.85	0.82	0.82	0.65	0.77
Recall	0.97	1	0.91	0.88	0.85	0.9	0.85	0.82
F1 score	0.93	0.94	0.87	0.86	0.83	0.86	0.74	0.79
Log loss	3.03	2.42	5.45	5.45	4.56	3.91	7.82	9.12
Without NCDs								
Accuracy	0.93	0.91	0.84	0.86	0.91	0.88	0.74	0.74
AUC	0.92	0.91	0.83	0.85	0.9	0.86	0.73	0.71
Precision	0.89	0.91	0.83	0.85	0.89	0.84	0.74	0.86
Recall	1	0.94	0.91	0.91	0.97	0.97	0.81	0.48
F1 score	0.94	0.92	0.87	0.88	0.93	0.9	0.78	0.62
Log loss	2.42	3.02	5.45	4.85	3.03	4.24	9.09	9.09

[a]RF: random forest.

[b]LGBM: light gradient boosting machine.

[c]SVM: support vector machine.

[d]DT: decision tree.

[e]XGB: extreme gradient boosting.

[f]GBM: gradient boosting machine.

[g]KNN: k-nearest neighbor.

[f]ANN: artificial neural network.

[g]AUC: area under the curve.

[h]NCDs: noncommunicable diseases.

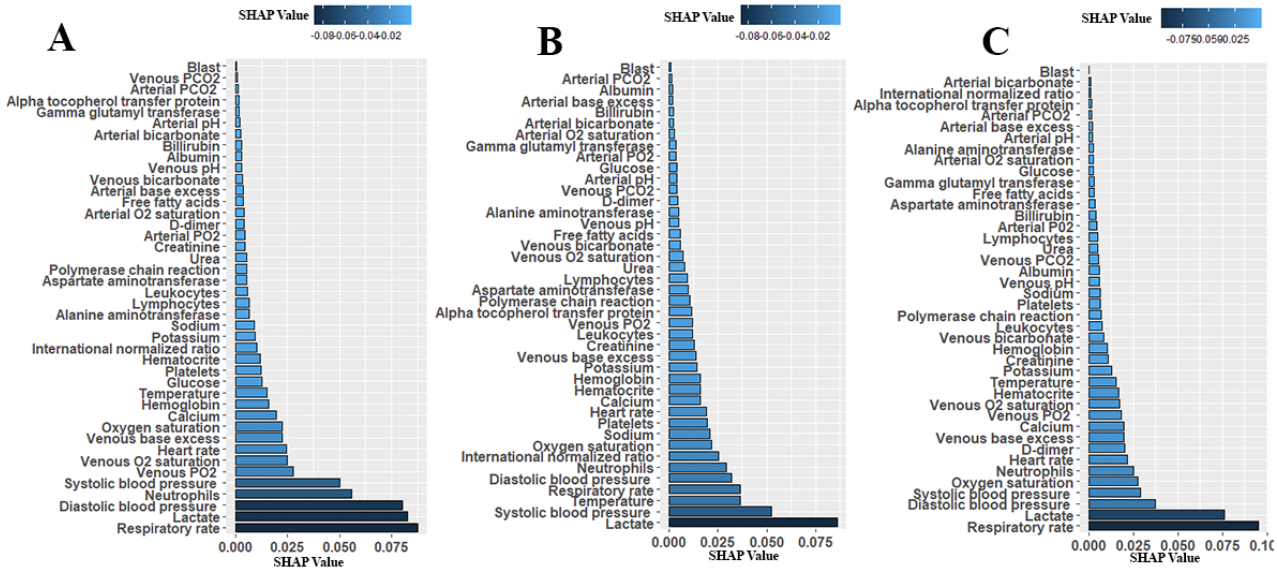
In the data group of patients with no NCDs, we found that RF demonstrated the highest accuracy score of 93%, LGBM and XGB performed with 91%, and SVM and DT showed good accuracy scores of >80%. However, KNN and ANN showed comparatively low accuracy scores of 74% because when we divided the data set, the size of the data was small. RF demonstrated the highest AUC of 92%; the AUC of LGBM was 91% and that of XGB was 90%. LGBM showed the highest precision value of 91%, while RF and XGB showed values of 89%. The highest precision value was 91% for LGBM, and other methods and models had values >80% except for KNN (74%). The highest recall values were 100% for RF and 97% for XGB and GBM; the other methods and models showed values above 80%, except ANN (48%). RF achieved the highest F1 score of 94%; XGB achieved a score of 93%, LGBM scored 92%, and SVM and DT scored 88%. However, KNN and ANN achieved comparatively low F1 scores, with 78% and 62% respectively, because of the lower training sample sizes. The lowest log loss value was 2.42% for RF, and the other methods and models also demonstrated good log loss values below 10%. In this patient group, we observed that excepting KNN and ANN, all of the models achieved accuracy scores >80%, and the evaluation matrix showed good model performance. Therefore, the best-performing models could be usefully applied in clinical scenarios.

In the data group of patients who had one or more coexisting NCDs, we found that LGBM performed with the highest accuracy score of 93%, and RF, GBM, XGB, SVM, and DT achieved scores of 91%, 89%, 87%, 84%, and 84%, respectively. KNN and ANN performed poorly, showing 77% and 74% accuracy, respectively; however, this result was due to the small

amount of available data. The highest AUC score was 92% for LGBM, and RF, SVM, DT, XGB, GBM, KNN, and ANN scored 91%, 83%, 84%, 87%, 89%, 79% and 71%, respectively. RF and LGBM demonstrated the highest precision value of 89%, and the other methods and models performed with good precision values >80%, except for KNN and ANN. LGBM achieved the highest recall value of 100%, RF achieved 97%, GBM 90%, SVM 83%, and DT 88%; the other methods and models performed above 80%. The highest F1 score was 94% for LGBM; RF also demonstrated 93%, and the other methods and models performed above 80% except for KNN and ANN. KNN and ANN achieved F1 scores of 74% and 79%, respectively; however, the number of training samples for these models was small.

Using ML analysis, we attempted to determine the most significant blood parameters that are highly predictive for identifying patients with severe COVID-19. We found the SHAP (Shapley Additive Explanations) values for each of the ML algorithms, quantile-normalized those values, and finally calculated the average values for each blood parameter. In Figure 6, the parameter list sorted according to the feature importance level (average SHAP value) is presented. In this figure, the left panel shows the combined patients (those with NCDs and those without NCDs), the middle panel shows the patients who have NCDs only, and the right panel shows the patients who have no NCDs.

Figure 6. Sorted significant and impacted blood parameters of patients with COVID-19 based on SHAP values, defined as the coefficient values of each parameter after model training: (A) combined patients group; (B) patients with noncommunicable diseases; (C) patients without noncommunicable diseases. Artificial intelligence models were used to identify the most predictive blood parameters for the severity of COVID-19 symptoms. Higher coefficient values of machine learning model outcomes indicate a higher significant association with disease severity. D-dimer: dimerized plasmin fragment D; FFA: free fatty acids; GGT: gamma-glutamyl transferase; INR: international normalized ratio; SHAP: Shapley Additive Explanations; TTPA: partial thromboplastin time.



In the above analysis, we observed that a small set of blood parameters had high SHAP values, which indicates that those parameters are impactful and predictable for the diagnosis of severe COVID-19. According to the level of importance, respiratory rate, lactate, blood pressure (diastolic and systolic), neutrophils, and oxygen saturation level were the most significant and common parameters for the group including all the patients. The exceptional cases are venous PO_2, venous saturated O_2, and heart rate, which were impactful for the combined patient group, and temperature and INR, which were impactful for the group of patients with NCDs only.

In the statistical analysis, it was found that the absolute value of lymphocytes is a key predictor for severe patient outcomes. The value of the lymphocytes parameter decreased with increasing severity level of the patients with COVID-19. We also observed the opposite scenario for neutrophil data, as in, the lymphocytes parameter increased if the patient's condition deteriorated toward a severe situation.

Discussion

Principal Findings

During the worldwide outbreak of COVID-19, classifications of disease mortality risk are of very great significance in prevention and treatment allocation. In this investigation, we identified a number of blood analysis parameters that can be used as risk factors for the assessment of disease severity in patients with COVID-19. We developed predictive algorithms that use a large number of blood parameters and demonstrated that these methods have potential to predict the disease severity of patients with COVID-19 with high accuracy.

We identified a number of features of patient data that contributed strongly to the predicted value of the algorithms (ie, were found to contribute to the accuracy of all our best ML algorithms), some of which were not obvious candidate predictors. We found that the absolute value of lymphocytes in the group of patients with severe symptoms was consistently lower than that in the nonsevere symptom group. The neutrophil

parameters of the severe symptom group were higher than those of the nonsevere symptom group. A high neutrophil level indicates a heightened level of immune activation and may play a role in the "inflammatory storm" that is characteristic of severe COVID-19 symptoms, which results in great harm to tissues and cells [28]. Low lymphocyte levels may reflect impeded antibody-based immune cell functions, which are suspected to result in patients with severe COVID-19 who are susceptible to bacterial infection [29]. Our results suggest that the numbers of circulating lymphocytes in the patients who developed severe symptoms were significantly lower than those in patients who did not have severe symptoms. In contrast, the inclusion of neutrophils in the severe patients in the ICU showed a greater influence, which is consistent with the findings of Qin et al [30].

We found that the indicator factors could be reliable predictors that discriminated between patients with severe and nonsevere COVID-19. Recent work has revealed the utility of routine blood parameters in the screening of patients with COVID-19. This is facilitated by the fact that blood parameter analysis is generally fast, affordable, and promptly accessible in the same health facility where patients are receiving treatment. The pathological tests of patients with COVID-19 identified abnormalities in some blood parameters. In previous published studies, a number of altered blood parameters in patients with COVID-19 who developed severe symptoms were identified in addition to the lymphocyte and neutrophil parameters noted above, such as eosinophils, basophils, monocytes, platelets, and total leukocytes as well as serum levels of urea, potassium, hemoglobin, and C-reactive blood protein [31-33]; this provides supportive evidence for our findings. Li et al [34] identified that bacterial infection affected COVID-19 pneumonia in some cases of mortality. Bacterial contamination also causes expanded leucocyte count and neutrophil count, which may be linked to defective immune responses. A few patients with COVID-19 have abnormal blood coagulation function: prothrombin time and D-dimer level increase [28], while thrombosis is linked with expanded platelet consumption and diminished platelet number.

Respiratory rate is one of the principal vital signs for symptom severity in patients with COVID-19. Abnormally high respiratory rates (<12 or >25 breaths/min) are also seen in a range of conditions, including asthma, heightened anxiety, pneumonia, congestive heart failure, and lung disease (all of which exacerbate COVID-19 conditions when presenting as comorbidities) and are a significant feature in severely affected patients with COVID-19 [35,36]. Elevated heart rate is similarly a key sign [37] and may be a cause of dizziness or shortness of breath in patients with sCOVID-19 [38]. Blood pressure is additionally a clinical sign for patients with COVID-19 [39]. Hypoxemia is also a sign that indicates a below-average level of oxygen saturation in the blood. The usual range of arterial oxygen is approximately 75-100 mm Hg, and a pulse oximeter reads the expected range from 95% to 100%; below 90% indicates that the patient's condition is critical [40]. This finding is often observed in patients with COVID-19 who may lack other obvious symptoms; therefore, it is a particularly dangerous feature of the disease. The serum lactic acid test is also a significant test that indicates disease severity in patients with COVID-19. Typically, the level of lactate in the blood is very low; a rise in lactate level is typically associated with low oxygen levels [41,42].

In summary, a number of signs and symptoms can indicate that COVID-19 is likely to become severe in a patient. A standardized and objective way to combine these and other less obvious predictors in a way that can optimize patient outcomes and resource management is needed. Our methodology, described here and derived from a number of different ML algorithms, can provide such an improved method. Indeed, the fact that high accuracy was obtained using similar predictors by different ML algorithms (indicating that there is limited sensitivity to the methodology) can provide confidence that these parameters are useful and that the approach is a sound one.

Conclusion

The results of our analysis indicated that there is a strong relationship between particular abnormal blood parameters and disease severity status in hospitalized patients with COVID-19. The primary utility of our findings is that the subset of routine blood parameters linked to disease severity could be used in a predictive algorithm that would better enable appropriate care to be given before the onset of severe symptoms. This is of particular importance in developing countries, where ICU beds in hospitals are a limited resource. This can be achieved using a relatively small number of currently available blood-based hospital tests to properly use ICU resources and identify patients who need to be monitored closely.

Among the association between blood parameters that can give predictive information regarding the severity of COVID-19 symptoms, the levels of lactate and immature granulocytes (absolute) appeared to have the strongest predictive value. Levels of hemoglobin, procalcitonin, erythrocyte sedimentation rate, brain natriuretic peptide, ferritin, D-dimer, and platelets likewise showed significant deviation from the normal control group for prediction of disease severity. Other parameters, namely respiratory rate, lactate, blood pressure (systolic and diastolic), hematocrit, venous and arterial base excess, neutrophils, albumin, and urea, showed less obvious deviations but clearly had predictive value. Our work suggests that links exist between these parameters and COVID-19, and similar proinflammatory infectious diseases may merit more detailed physiological investigations.

There were a few limitations to our study. First, the small sample size may restrict the precision of the identification of severity. Second, the absence of more detailed clinical information in the data sets that were used (such as patient age, sex, and comorbidities) may hinder better classification, although this suggests that in future studies, we could use new data sets to address this and improve on our work. Finally, the disease severity and mortality of COVID-19 varies significantly from country to country; the reasons for this are very poorly understood, but it is suggested that this type of predictive analysis should be conducted on data from other parts of the world to improve the performance of the algorithm. Nevertheless, we hope our study can be used by practitioners

and help policy makers to improve resource allocation and outcomes for patients with COVID-19.

Acknowledgments

This research was supported by the Deanship of Scientific Research, Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia (Grant No. 21-13-18-008).

Conflicts of Interest

None declared.

References

1. Mohammadi M, Meskini M, do Nascimento Pinto AL. 2019 Novel coronavirus (COVID-19) overview. Z Gesundh Wiss 2020 Apr 19:1-9 [FREE Full text] [doi: 10.1007/s10389-020-01258-3] [Medline: 32313806]
2. Yang J, Chen X, Deng X, Chen Z, Gong H, Yan H, et al. Disease burden and clinical severity of the first pandemic wave of COVID-19 in Wuhan, China. Nat Commun 2020 Oct 27;11(1):5411 [FREE Full text] [doi: 10.1038/s41467-020-19238-2] [Medline: 33110070]
3. Ahamad MM, Aktar S, Rashed-Al-Mahfuz M, Uddin S, Liò P, Xu H, et al. A machine learning model to identify early stage symptoms of SARS-Cov-2 infected patients. Expert Syst Appl 2020 Dec 01;160:113661 [FREE Full text] [doi: 10.1016/j.eswa.2020.113661] [Medline: 32834556]
4. Nashiry A, Sarmin Sumi S, Islam S, Quinn J, Moni M. Bioinformatics and system biology approach to identify the influences of COVID-19 on cardiovascular and hypertensive comorbidities. Brief Bioinform 2021 Mar 22;22(2):1387-1401 [FREE Full text] [doi: 10.1093/bib/bbaa426] [Medline: 33458761]
5. Taz T, Ahmed K, Paul B, Al-Zahrani F, Mahmud S, Moni M. Identification of biomarkers and pathways for the SARS-CoV-2 infections that make complexities in pulmonary arterial hypertension patients. Brief Bioinform 2021 Mar 22;22(2):1451-1465 [FREE Full text] [doi: 10.1093/bib/bbab026] [Medline: 33611340]
6. Prin M, Wunsch H. International comparisons of intensive care. Curr Opin Crit Care 2012;18(6):700-706. [doi: 10.1097/mcc.0b013e32835914d5]
7. Satu M, Khan M, Rahman M, Howlader KC, Roy S, Roy SS, et al. Diseasome and comorbidities complexities of SARS-CoV-2 infection with common malignant diseases. Brief Bioinform 2021 Mar 22;22(2):1415-1429 [FREE Full text] [doi: 10.1093/bib/bbab003] [Medline: 33539530]
8. Uddin S, Imam T, Ali MM. The implementation of public health and economic measures during the first wave of COVID-19 by different countries with respect to time, infection rate and death rate. 2021 Feb Presented at: 2021 Australasian Computer Science Week Multiconference; February 1-5, 2021; Online conference p. 1-8. [doi: 10.1145/3437378.3437384]
9. Hong KH, Lee SW, Kim TS, Huh HJ, Lee J, Kim SY, et al. Guidelines for laboratory diagnosis of coronavirus disease 2019 (COVID-19) in Korea. Ann Lab Med 2020 Sep 01;40(5):351-360 [FREE Full text] [doi: 10.3343/alm.2020.40.5.351] [Medline: 32237288]
10. Nain Z, Rana H, Liò P, Islam S, Summers M, Moni M. Pathogenetic profiling of COVID-19 and SARS-like viruses. Brief Bioinform 2021 Mar 22;22(2):1175-1196 [FREE Full text] [doi: 10.1093/bib/bbaa173] [Medline: 32778874]
11. Taz T, Ahmed K, Paul B, Kawsar M, Aktar N, Mahmud SMH, et al. Network-based identification genetic effect of SARS-CoV-2 infections to Idiopathic pulmonary fibrosis (IPF) patients. Brief Bioinform 2021 Mar 22;22(2):1254-1266 [FREE Full text] [doi: 10.1093/bib/bbaa235] [Medline: 33024988]
12. Li Z, Yi Y, Luo X, Xiong N, Liu Y, Li S, et al. Development and clinical application of a rapid IgM-IgG combined antibody test for SARS-CoV-2 infection diagnosis. J Med Virol 2020 Sep 13;92(9):1518-1524 [FREE Full text] [doi: 10.1002/jmv.25727] [Medline: 32104917]
13. Stachel A. Development and validation of a machine learning model for use as an automated artificial intelligence tool to predict mortality risk in patients with COVID-19. Zenodo. 2020 Jun 14. URL: http://doi.org/10.5281/zenodo.3893846 [accessed 2020-11-16]
14. COVID-19 - clinical data to assess diagnosis. Kaggle. 2020 Jun 22. URL: https://www.kaggle.com/S%C3%ADrio-Libanes/covid19 [accessed 2020-11-16]
15. Nihan ST. Karl Pearsons chi-square tests. Educ Res Rev 2020 Sep 30;15(9):575-580 [FREE Full text] [doi: 10.5897/ERR2019.3817]
16. Horne A. Statistics, use in immunology. In: Encyclopedia of Immunology. Amsterdam, Netherlands: Elsevier; 1998:2211-2215.
17. 11. Correlation and regression. The BMJ. URL: https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression [accessed 2020-11-16]
18. Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. J Comput Sci Eng 2018 Oct 31;6(10):74-78. [doi: 10.26438/ijcse/v6i10.7478]

19. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 2019 Dec 21;19(1):281 [FREE Full text] [doi: 10.1186/s12911-019-1004-8] [Medline: 31864346]
20. Aluja-Banet T, Nafria E. Stability and scalability in decision trees. Comput Stat 2015 Feb 26;18(3-4):505-520. [doi: 10.1007/bf03354613]
21. Sciabola S, Fang C. Gradient boosting decision tree models for better temporal ADME prediction from an industrial perspective. 2020 Aug 19 Presented at: ACS Fall 2020 Virtual Meeting; August 17-20, 2020; virtual meeting. [doi: 10.1021/scimeetings.0c06777]
22. Hutter F, Hoos H, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: Lecture Notes in Computer Science Learning and Intelligent Optimization. 2011 Presented at: LION 2011: International Conference on Learning and Intelligent Optimization; May 24-28, 2020; Athens, Greece p. 507-523. [doi: 10.1007/978-3-642-25566-3_40]
23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17: the 31st International Conference on Neural Information Processing Systems; Long Beach, CA; December 4-9, 2017 p. 4768-4777.
24. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan 17;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
25. Wang R, Li J. Bayes test of precision, recall, and F1 measure for comparison of two natural language processing models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics; July 2019; Florence, Italy. [doi: 10.18653/v1/p19-1405]
26. Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, et al. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. J Clin Epidemiol 2020 Oct;126:207-216. [doi: 10.1016/j.jclinepi.2020.01.028] [Medline: 32712176]
27. Kiapour A. Bayes, E-Bayes and robust Bayes premium estimation and prediction under the squared log error loss function. JIRSS 2018 Jun 01;17(1):33-47. [doi: 10.29252/jirss.17.1.33]
28. Mo P, Xing Y, Xiao Y, Deng L, Zhao Q, Wang H, et al. Clinical characteristics of refractory COVID-19 pneumonia in Wuhan, China. Clin Infect Dis 2020 Mar 16:2020 [FREE Full text] [doi: 10.1093/cid/ciaa270] [Medline: 32173725]
29. 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]
30. Qin C, Zhou L, Hu Z, Zhang S, Yang S, Tao Y, et al. Dysregulation of Immune Response in Patients with COVID-19 in Wuhan, China. SSRN Journal. Preprint posted online on March 2, 2020 2020. [doi: 10.2139/ssrn.3541136]
31. AlJame M, Ahmad I, Imtiaz A, Mohammed A. Ensemble learning model for diagnosing COVID-19 from routine blood tests. Inform Med Unlocked 2020;21:100449 [FREE Full text] [doi: 10.1016/j.imu.2020.100449] [Medline: 33102686]
32. Li X, Wang L, Yan S, Yang F, Xiang L, Zhu J, et al. Clinical characteristics of 25 death cases with COVID-19: a retrospective review of medical records in a single medical center, Wuhan, China. Int J Infect Dis 2020 May;94:128-132 [FREE Full text] [doi: 10.1016/j.ijid.2020.03.053] [Medline: 32251805]
33. Sun S, Cai X, Wang H, He G, Lin Y, Lu B, et al. Abnormalities of peripheral blood system in patients with COVID-19 in Wenzhou, China. Clin Chim Acta 2020 Aug;507:174-180 [FREE Full text] [doi: 10.1016/j.cca.2020.04.024] [Medline: 32339487]
34. Li X, Wang L, Yan S, Yang F, Xiang L, Zhu J, et al. Clinical characteristics of 25 death cases with COVID-19: a retrospective review of medical records in a single medical center, Wuhan, China. Int J Infect Dis 2020 May;94:128-132 [FREE Full text] [doi: 10.1016/j.ijid.2020.03.053] [Medline: 32251805]
35. Bernardi L, Porta C, Gabutti A, Spicuzza L, Sleight P. Modulatory effects of respiration. Autonomic Neuroscience 2001 Jul;90(1-2):47-56. [doi: 10.1016/s1566-0702(01)00267-3]
36. Lee M. Clinical characteristics of early noncritical hospitalized patients with coronavirus disease. 2020 Presented at: 1st Annual Mount Sinai Morningside and Mount Sinai West Internal Medicine Residency Program's Research Week; May 26-29, 2020; New York, NY. [doi: 10.26226/morressier.5ebc261fffea6f735881a237]
37. Peer N, Lombard C, Steyn K, Levitt N. Elevated resting heart rate is associated with several cardiovascular disease risk factors in urban-dwelling black South Africans. Sci Rep 2020 Mar 12;10(1):4605 [FREE Full text] [doi: 10.1038/s41598-020-61502-4] [Medline: 32165685]
38. Pavri BB, Kloo J, Farzad D, Riley JM. Behavior of the PR interval with increasing heart rate in patients with COVID-19. Heart Rhythm 2020 Sep;17(9):1434-1438 [FREE Full text] [doi: 10.1016/j.hrthm.2020.06.009] [Medline: 32535142]
39. Lazic S, Lazic B. The correlation between systolic and diastolic blood pressure and diastolic parameters in arterial hypertension in the presence of normal systolic function. Cardiol Croat 2014 May 22;9(5-6):166-166. [doi: 10.15836/ccar.2014.166]
40. Anusha B, Madhusudhana K, Chinni SK, Paramesh Y. Assessment of pulp oxygen saturation levels by pulse oximetry for pulpal diseases –a diagnostic study. J Clin Diagn Res 2017 Sep;11(9):ZC36-ZC39. [doi: 10.7860/jcdr/2017/28322.10572]

41. Aktar S, Talukder A, Talukder A, Martuza Ahamad M, Kamal AHM, Khan JR, et al. Machine learning and meta-analysis approach to identify patient comorbidities and symptoms that increased risk of mortality in COVID-19. ArXiv. Preprint posted online on August 25, 2020 2020.
42. Tan L, Kang X, Ji X, Li G, Wang Q, Li Y, et al. Validation of predictors of disease severity and outcomes in COVID-19 patients: a descriptive and retrospective study. Med (N Y) 2020 Dec 18;1(1):128-138.e3 [FREE Full text] [doi: 10.1016/j.medj.2020.05.002] [Medline: 32838352]

Abbreviations

ANN: artificial neural network
AUC-ROC: area under the receiver operator characteristic curve
D-dimer: dimerized plasmin fragment D
DT: decision tree
GBM: gradient boosting machine
ML: machine learning
NCD: noncommunicable disease
ICU: intensive care unit
INR: international normalized ratio
KNN: k-nearest neighbor
LFC: log 2 fold change
LGBM: light gradient boosting machine
RF: random forest
RT-PCR: reverse transcription–polymerase chain reaction
SHAP: Shapley Additive Explanation
SVM: support vector machine
XGBoost: extreme gradient boosting

Edited by C Lovis; submitted 20.11.20; peer-reviewed by W Jiang, S Kriventsov; comments to author 23.12.20; revised version received 21.01.21; accepted 21.03.21; published 13.04.21.

Please cite as:
Aktar S, Ahamad MM, Rashed-Al-Mahfuz M, Azad AKM, Uddin S, Kamal AHM, Alyami SA, Lin PI, Islam SMS, Quinn JMW, Eapen V, Moni MA
Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development
JMIR Med Inform 2021;9(4):e25884
URL: https://medinform.jmir.org/2021/4/e25884
doi:10.2196/25884
PMID:33779565

Original Paper

Predicting Intensive Care Transfers and Other Unforeseen Events: Analytic Model Validation Study and Comparison to Existing Methods

Brandon C Cummings[1*], MHI; Sardar Ansari[1*], PhD; Jonathan R Motyka[1], MSc; Guan Wang[1], MSc; Richard P Medlin Jr[1], MSIS, MD; Steven L Kronick[1], MSc, MD; Karandeep Singh[1,2,3,4], MMSc, MD; Pauline K Park[1,5], MD; Lena M Napolitano[1,5], MD; Robert P Dickson[1,2,6], MD; Michael R Mathis[1,7], MD; Michael W Sjoding[1,2], MD; Andrew J Admon[1,2,4], MPH, MSc, MD; Ross Blank[1,7], MD; Jakob I McSparron[1,2], MD; Kevin R Ward[1,4,8], MD; Christopher E Gillies[1,4], PhD

[1]Michigan Center for Integrative Research in Critical Care, Department of Emergency Medicine, University of Michigan, Ann Arbor, MI, United States
[2]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States
[3]Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI, United States
[4]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI, United States
[5]Department of Surgery, University of Michigan, Ann Arbor, MI, United States
[6]Department of Microbiology & Immunology, University of Michigan, Ann Arbor, MI, United States
[7]Department of Anesthesiology, University of Michigan, Ann Arbor, MI, United States
[8]Department of Biomedical Engineering, University of Michigan, Ann Arbor, MI, United States
[*]these authors contributed equally

Corresponding Author:
Brandon C Cummings, MHI
Michigan Center for Integrative Research In Critical Care
Department of Emergency Medicine
University of Michigan
2800 N Plymouth Road
NCRC 10-A112
Ann Arbor, MI, 48109
United States
Phone: 1 (734) 647 7436
Email: cummingb@med.umich.edu

Abstract

Background: COVID-19 has led to an unprecedented strain on health care facilities across the United States. Accurately identifying patients at an increased risk of deterioration may help hospitals manage their resources while improving the quality of patient care. Here, we present the results of an analytical model, Predicting Intensive Care Transfers and Other Unforeseen Events (PICTURE), to identify patients at high risk for imminent intensive care unit transfer, respiratory failure, or death, with the intention to improve the prediction of deterioration due to COVID-19.

Objective: This study aims to validate the PICTURE model's ability to predict unexpected deterioration in general ward and COVID-19 patients, and to compare its performance with the Epic Deterioration Index (EDI), an existing model that has recently been assessed for use in patients with COVID-19.

Methods: The PICTURE model was trained and validated on a cohort of hospitalized non–COVID-19 patients using electronic health record data from 2014 to 2018. It was then applied to two holdout test sets: non–COVID-19 patients from 2019 and patients testing positive for COVID-19 in 2020. PICTURE results were aligned to EDI and NEWS scores for head-to-head comparison via area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve. We compared the models' ability to predict an adverse event (defined as intensive care unit transfer, mechanical ventilation use, or death). Shapley values were used to provide explanations for PICTURE predictions.

Results: In non–COVID-19 general ward patients, PICTURE achieved an AUROC of 0.819 (95% CI 0.805-0.834) per observation, compared to the EDI's AUROC of 0.763 (95% CI 0.746-0.781; n=21,740; *P*<.001). In patients testing positive for COVID-19,

PICTURE again outperformed the EDI with an AUROC of 0.849 (95% CI 0.820-0.878) compared to the EDI's AUROC of 0.803 (95% CI 0.772-0.838; n=607; *P*<.001). The most important variables influencing PICTURE predictions in the COVID-19 cohort were a rapid respiratory rate, a high level of oxygen support, low oxygen saturation, and impaired mental status (Glasgow Coma Scale).

Conclusions: The PICTURE model is more accurate in predicting adverse patient outcomes for both general ward patients and COVID-19 positive patients in our cohorts compared to the EDI. The ability to consistently anticipate these events may be especially valuable when considering potential incipient waves of COVID-19 infections. The generalizability of the model will require testing in other health care systems for validation.

(JMIR Med Inform 2021;9(4):e25066) doi:10.2196/25066

KEYWORDS

COVID-19; biomedical informatics; critical care; machine learning; deterioration; predictive analytics; informatics; prediction; intensive care unit; ICU; mortality

Introduction

The effect of COVID-19 on the US health care system is difficult to overstate. It has led to unprecedented clinical strain in hospitals nationwide, prompting the proliferation of intensive care unit (ICU) capability and of lower-acuity field hospitals to accommodate the increased patient load. A predictive early warning system capable of identifying patients at increased risk of deterioration could assist hospitals in maintaining a high level of patient care while more efficiently distributing their thinly stretched resources. However, a recent review has illustrated that high quality validated models of deterioration in patients with COVID-19 are lacking [1]. All 16 of the models appraised in this review were rated at high or unclear risk of bias, mostly because of nonrepresentative selection of control patients. A primary concern is that these models may overfit to the small COVID-19 data sets that are currently available.

Early warning systems have been and continue to be applied in hospital settings prior to the COVID-19 pandemic to predict patient deterioration events before they occur, giving health care providers time to intervene [2]. The prediction of adverse events such as ICU admission and death provides crucial information to avert impending critical deterioration; it is estimated that 85% of such events are preceded by detectable changes in physiological signs [3] that may occur up to 48 hours before the event [4]. In addition, approximately 44% of events are avoidable through early intervention [5], and 90% of unplanned transfers to the ICU are preceded by a new or worsening condition [6,7]. Such abnormal signals indicate that predictive data analytics may be used to alert providers of incipient deterioration events, ultimately leading to improved care and reduced costs [8,9]. Given the number of unknowns surrounding the pathophysiology of COVID-19, early warning systems may play a pivotal role in treating patients and improving outcomes.

One model that has been assessed in patients with COVID-19 is the Epic Deterioration Index (EDI; Epic Systems Inc) [10,11]. The EDI is a proprietary clinical early warning system that aims to identify patients at an increased risk of deterioration and who may require a higher level of care. The EDI has the advantage over models built on COVID-19–specific data in that it is not overfit to small data sets, as it was trained on over 130,000 encounters [11,12]. Recent work has suggested it may be capable of stratifying patients with COVID-19 according to their risk of deterioration [11]. The outcomes used in this study were those considered most relevant to the care of patients with COVID-19 including ICU level of care, mechanical ventilation, and death. Although the EDI was able to successfully isolate groups of patients at very high and very low risk of deterioration, the overall performance as a continuous predictor was moderately low (area under the receiver operating characteristic curve [AUROC] 0.76, 95% CI 0.68-0.84; n=174) [11]. Additionally, much of the detail surrounding the EDI's structure and internal validation has not been shared publicly. This makes the interpretation of individual predictions difficult. Since hospitals who do not use Epic electronic health record (EHR) systems may not have access to EDI predictions, we have also evaluated the publicly available National Early Warning Score (NEWS) as a secondary comparison.

In this study, we have applied our previously described model, Predicting Intensive Care Transfers and Other Unforeseen Events (PICTURE), to a cohort of patients testing positive for COVID-19 [13]. Initially developed to predict patient deterioration in the general wards, we have retrained the model to target those outcomes considered most relevant to the COVID-19 pandemic: ICU level of care, mechanical ventilation, and death. PICTURE, like the EDI, was trained and tuned on a large non–COVID-19 cohort including patients both with and without infectious diseases (131,546 encounters). Furthermore, we took extensive steps in the PICTURE framework to limit overfitting and learning missingness patterns in the data, such as a novel imputation mechanism [13]. This is critical in providing clinicians with novel, useful, and generalizable alerts, as missing patterns can vary in different settings and different patient phenotypes [13]. In addition to the risk score, PICTURE also provides actionable explanations for its predictions in the form of Shapley values, which may help clinicians easily interpret scores and better determine if actionability on the alert is required [14]. We validated this system in both a non–COVID-19 cohort and in patients who were hospitalized testing positive for COVID-19 and compared it to the EDI and NEWS on the same matched cohorts.

Methods

Setting and Study Population

The study protocol was approved by the University of Michigan's Institutional Review Board (HUM00092309). EHR data was collected from a large tertiary, academic medical system (Michigan Medicine) from January 1, 2014, to November 11, 2020. The first 5 years of data (2014-2018; n=131,546 encounters) were used to train and validate the model, while 2019 data was reserved as a holdout test set (n=33,472 encounters). Training, validation, and test populations were segmented to prevent overlap of multiple hospital encounters between sets. Criteria for inclusion in these three cohorts were defined as 18 years or older and who were hospitalized (having inpatient or other observation status) in a general ward. We excluded patients who were discharged to hospice and whose ICU transfer was from a floor other than a general ward (eg, operating or interventional radiology unit) to exclude planned ICU transfers. We also excluded patients with a left ventricular assist device to avoid artifactual blood pressure readings.

To be included in the COVID-19 cohort (n=637 encounters), patients must have been admitted to the hospital with a COVID-19 diagnosis and have received a positive COVID-19 test from Michigan Medicine during their encounter. These patients were then filtered using the same criteria used in the 2019 test set, with the exception of the hospice distinction. Only discharged patients or those who already experienced an adverse event were included. Table 1 describes the study cohort and the frequency of individual adverse events. When compared to the non–COVID-19 test cohort from 2019, the proportion of Black and Asian patients was significantly higher (Black: 4214/33,472, 12.6% vs 220/637, 34.5%; *P*<.001; Asian: 686/33,472, 2.0% vs 29/637, 4.6%; *P*<.001). The rate of adverse events was also higher, rising from 4.0% (1337/33,472) to 24.3% (155/637; *P*<.001).

Table 1. Study population.[a]

Data set	Non–COVID-19			COVID-19	*P* value (non–COVID-19 vs COVID-19 test sets)[b]
	Training 2014-2018	Validation 2014-2018	Testing 2019	Testing 2020	
Encounters, n	105,457	26,089	33,472	637	N/A[c]
Patients, n	62,392	15,597	23,368	600	N/A
Age (years), median (IQR)	60.2 (46.5-70.8)	60.4 (46.7-71.2)	61.0 (47.0-71.5)	61.8 (49.6-72.0)	.02
Race, n (%)					
White	86,522 (82.0)	21,647 (83.0)	27,036 (80.8)	329 (51.6)	<.001
Black	12,344 (11.7)	2861 (11.0)	4214 (12.6)	220 (34.5)	<.001
Asian	2145 (2.0)	504 (1.9)	686 (2.0)	29 (4.6)	<.001
Other[d]	4446 (4.2)	1077 (4.1)	1536 (4.6)	59 (9.3)	<.001
Female sex, n (%)	53,225 (50.5)	13,048 (50.0)	16,760 (50.1)	282 (44.3)	.003
Event rate[e], n (%)	4236 (4.0)	1007 (3.9)	1337 (4.0)	155 (24.3)	<.001
Death	920 (0.9)	232 (0.9)	277 (0.8)	16 (2.5)	<.001
ICU[f] transfer	2979 (2.8)	717 (2.7)	1000 (3.0)	139 (21.8)	<.001
Mechanical ventilation	1330 (1.3)	299 (1.1)	352 (1.1)	49 (7.7)	<.001
Cardiac arrest[g]	143 (0.1)	37 (0.1)	56 (0.2)	N/A	N/A

[a]Patients were subset into one of four study cohorts: a training set for learning model parameters, a validation set for model structure and hyperparameter tuning, a holdout test set for evaluation, and a final test set composed of patients testing positive for COVID-19. Values are based on individual hospital encounters.

[b]*P* values were calculated across the two test sets using a Mann-Whitney *U* test for continuous variables (age) and a chi-square test for categorical variables.

[c]N/A: not applicable.

[d]Other races comprising less than 1% of the population each were incorporated under the "Other" heading.

[e]The event rate represents a composite outcome indicating that one of the following events occurred: death, ICU transfer, mechanical ventilation, and cardiac arrest. The individual frequencies of these adverse events are also reported and represent the number of cases where each particular outcome was the first to occur. Please see the section Outcomes for the procedure of calculating these targets.

[f]ICU: intensive care unit.

[g]Cardiac arrest was not used as a target in the COVID-19 positive population, as the manually adjudicated data is not yet available at the time of writing.

Predictors

The variables used as predictors were collected from the EHR and broadly included vital signs and physiologic observations, laboratory and metabolic values, and demographics. We selected specific features based on previous analysis [13]. Vital signs used in the model included heart rate, respiratory rate, pulse oximetry, Glasgow Coma Scale (GCS), urine output, and blood pressure. Laboratory and metabolic features included electrolyte concentrations, glucose and lactate, and blood cell counts. Demographics included age, height, weight, race, and gender. Fluid bolus and oxygen supplementation were also included as features. A full list of features is presented in Table S1 in Multimedia Appendix 1 alongside their respective median, IQR, and missingness rate. Variables centered on treatment (eg, medication administration) were largely excluded as, similar to the missingness flags described in Gillies et al [13], the scores generated by the model may be less generalizable and novel to the clinician as patterns of care change between diseases (eg, COVID-19) or institutions. Multimedia Appendix 1 Table S2 describes the effects of including medications as features in more detail.

Outcomes

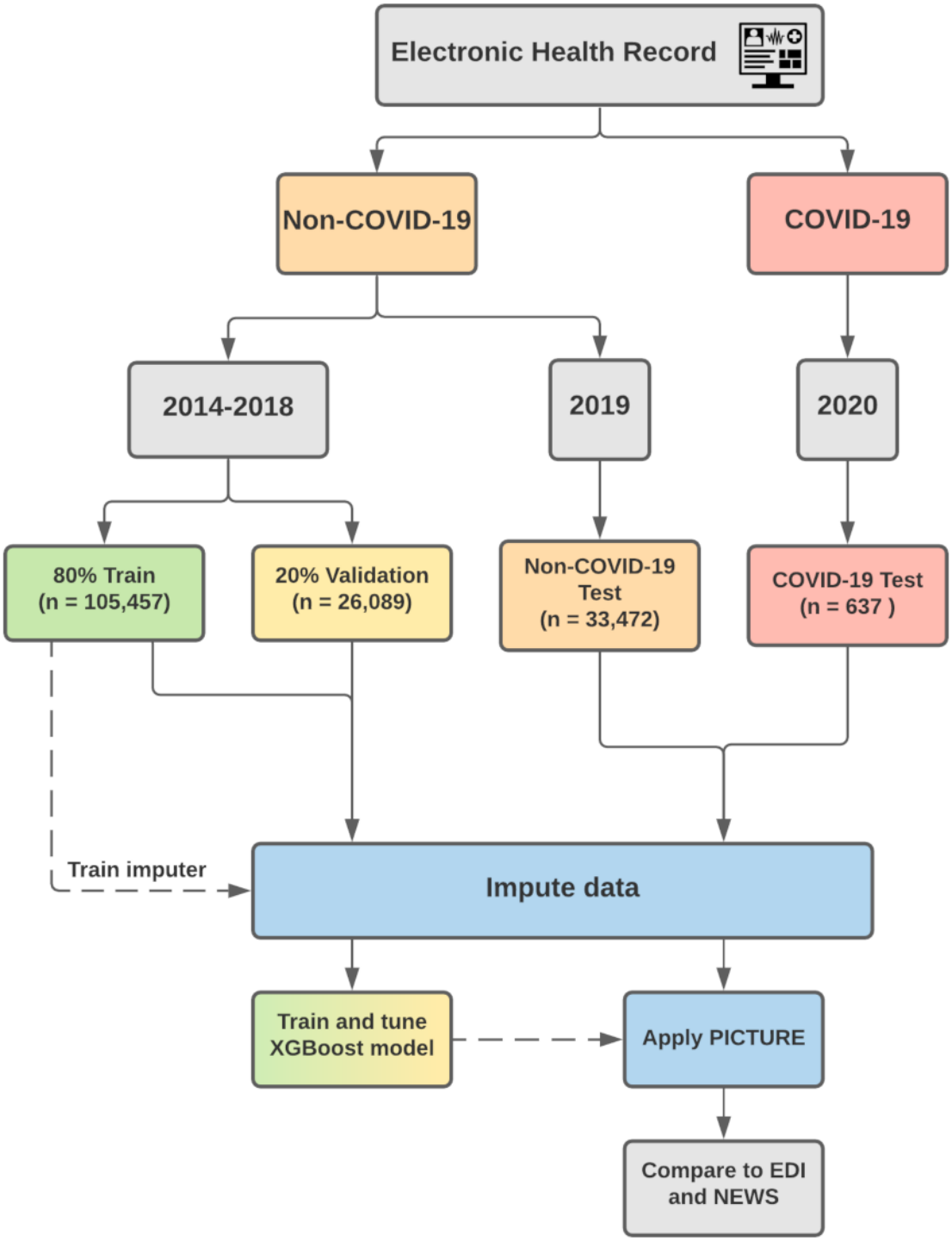
The primary outcomes in the training, validation, and non–COVID-19 test cohorts (data collected from 2014 through 2019) were death, cardiac arrest (as defined by the American Heart Association's *Get With The Guidelines*), transfer to an ICU from a general ward or similar unit, or need for mechanical ventilation. Determination of ICU transfer was based on actual location or accommodation level. Outcomes in the COVID-19 positive cohort differed slightly in two respects. First, cardiac arrest information was not available at the time of writing and so was not included. Second, the emergency procedures undertaken by the hospital to accommodate the high volume of patients with COVID-19 led to the delivery of critical care in non-ICU settings. Thus, "ICU level of care" is used to denote patients who were treated by ICU staff or given ICU-level care but who may not have been physically housed in a bed previously demarcated as an ICU bed. This information is derived from the admission, discharge, and transfer table. Level of care was used to determine ICU transfer in patients with COVID-19 in addition to actual location. We discarded observations occurring 30 minutes before the first event or later to be consistent with other approaches [15]. For observation-level predictions, individual observations were labeled positive if they occurred within 24 hours of any of the aforementioned events and negative otherwise. We refer to these composite adverse events as the *outcome* or *target* throughout the text. These outcomes were designed to closely follow those of a recent analysis of the EDI at Michigan Medicine [11].

To verify the accuracy of our automatically generated labels, a clinician (author MRM) manually reviewed the patient charts for 20 encounters to determine whether the patient was infected with COVID-19, whether the recorded event truly took place, and whether the event was unplanned. To do so, we randomly sampled two encounters (one positive, the other negative if available) from each patient service with eight or more encounters to ensure the accuracy of the labels across all services. The result was a sample of 20 encounters, 11 of which were positive. The recorded event of interest for each encounter was reviewed by the clinician to determine whether the event took place and whether it was emergent (not planned). For the patients that were labeled as negative, the clinician reviewed the entire patient chart to ensure that no adverse events occurred during the encounter. The results indicate that all 20 patients were infected with COVID-19, all the labels and the event times were accurate, and all the events were unplanned. This provides evidence that the automatically generated outcomes accurately identify unplanned adverse events.

PICTURE Model Development

To train and evaluate the PICTURE model, we partitioned our data into four folds: a training and validation set using data from 2014 to 2018, a test set using 2019 data, and a fourth set consisting of data from patients who are COVID-19 positive. We partitioned the sets such that multiple hospital encounters from the same individual were restricted to one cohort, preventing patient-level overlap between cohorts. Encounters with an admission date from January 1, 2014, to December 31, 2018, were used for training and validation and hyperparameter tuning (n=131,546 encounters). These patients were further divided between training and validation sets using an 80%/20% split. Those patients with an admission date between January 1 and December 31, 2019, were reserved as a holdout test set (n=33,472 encounters). Lastly, patients testing positive for COVID-19 from March 1 to September 11, 2020, were reserved as a separate set (n=637 encounters). Figure 1 displays a graphical overview of this delineation.

Figure 1. PICTURE training and validation framework. The electronic health record data is split into COVID-19 and non–COVID-19 patients. Encounters with an admission date between January 1, 2014, and December 31, 2018, were set aside for training (80%) and validation (20%) subsets. Encounters with an admission date between January 1 and December 31, 2019, were used as a non–COVID-19 test set. Encounters from 2020 that tested positive for COVID-19 were held out as a separate test set. In the case that a given patient has multiple encounters that overlap these boundaries, only the later encounters were considered to remove patient overlap between the cohorts. EDI: Epic Deterioration Index; NEWS: National Early Warning Score; PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events; XGBoost: extreme gradient boosting.



As the EHR stores data in a long format (with each new row corresponding to a new measurement at a new time point), it was first converted to a wide structure such that each observation represented all features at a given time point for a given patient. The training and validation sets were grouped into 8-hour windows to ensure that each encounter would have the same amount of observations for the same amount of time in the hospital, avoiding emphasis on patients who get more frequent updates while training the model as described in Gillies et al [13]. The 2019 and COVID-19 test sets were left in a granular

format, where each new observation represented the addition of new data (eg, an updated vital sign). Vital signs and laboratory values were forward filled such that each observation represented the most up-to-date information available as of that time, and the only time series–adjusted variables were oxygen supplementation, oxygen device use, and oxygen saturation as measured by pulse oximetry (SpO_2), which were represented by the maximum (oxygen supplementation and device) or minimum (SpO_2) over the previous 24 hours. Otherwise, each observation contained only the most up-to-date data available as of that time point and did not take historical values in to account. The remaining missing values were iteratively imputed using the mean of the posterior distribution from a multivariate Bayesian regression model. This method has previously been demonstrated to reduce the degree to which tree-based models learn missingness patterns to bolster performance [13]. Classification was achieved using an extreme gradient boosting model (v 0.90), an open-source implementation of a gradient-boosting tree framework that fits additional iterations using the errors of previous results [16]. The model uses a binary cross-entropy objective function with a maximum tree depth of three nodes, a learning rate of 0.05, no minimum loss reduction, uniform sampling with a subsample parameter of 0.6, and stopped when the validation area under the precision-recall curve (AUPRC) had not improved for 30 rounds. The model was applied to individual observations independently—that is, the model used the latest information available (via forward filling). In this sense, time dependance was not modeled aside from those aforementioned variables. All analyses were performed using Python 3.8.2 (Python Software Foundation).

Epic Deterioration Index and NEWS

The EDI is a proprietary model developed by Epic Systems Corporation. Michigan Medicine uses Epic as its electronic medical record system and has access to the EDI tool. Similar to PICTURE, it uses clinical data that are commonly available in the EHR to make predictions regarding patient deterioration. It was trained using a similar composite outcome including death, ICU transfer, and resuscitation as adverse events [11]. It is calculated every 15 minutes. Specific details surrounding its structure, parameters, or training procedures have not been shared publicly.

NEWS, developed by the Royal College of Physicians, is a second index used to detect patients at an increased risk of deterioration event such as cardiac arrest, ICU transfer, and death [17,18]. In contrast to the EDI, which is based on a proprietary system, the basis of the NEWS score is openly available. NEWS scores were calculated based on the algorithm described in Smith et al [17]. The original NEWS was selected over the updated NEWS2 score due to evidence that its performance was found to be higher when predicting adverse events in patients at risk of respiratory failure [19].

PICTURE Model Evaluation

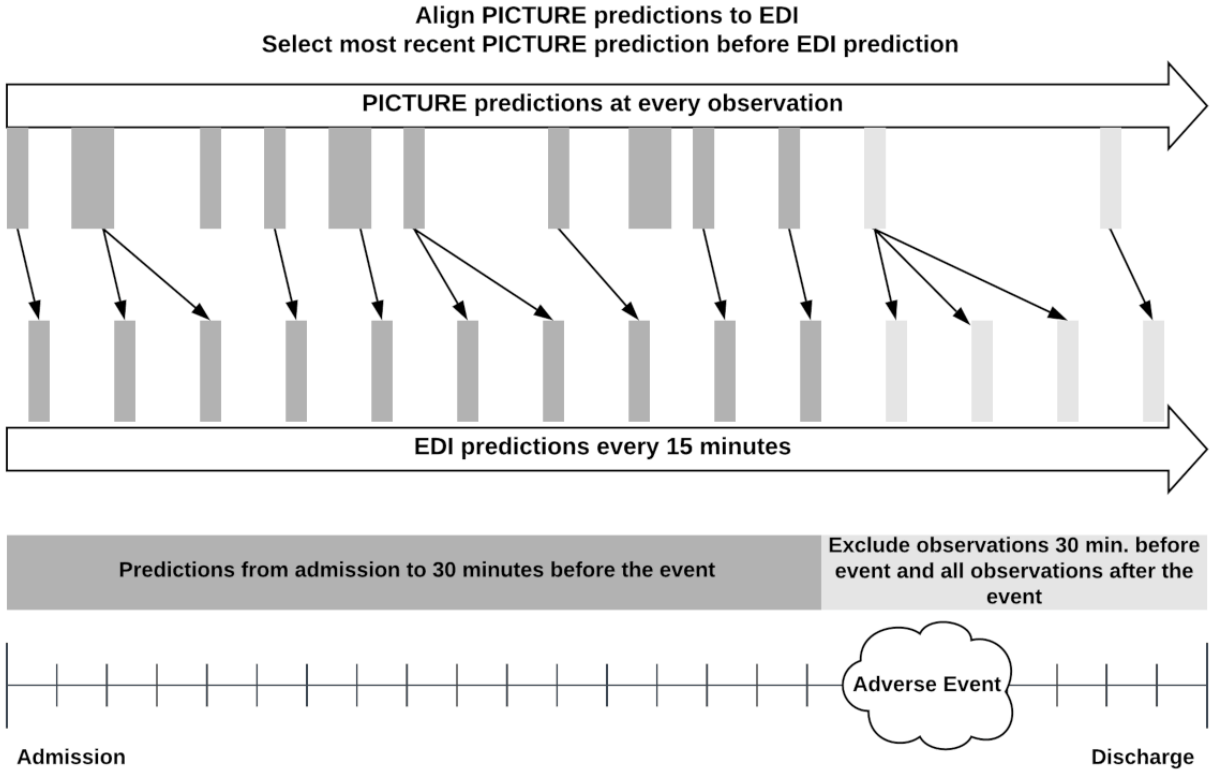
Evaluation of PICTURE Performance in Non–COVID-19 Cohort

We first assessed the performance of the PICTURE model on all 33,472 encounters in the holdout test set comprising patients from 2019. Another early warning aggregate score, NEWS, was used for comparison in this preliminary analysis [17,18]. For each observation time point, the NEWS score was calculated according to their published scoring system and compared to PICTURE scores. Performance was assessed on two scales: observation level and encounter level. The term *observation level* is used to denote the performance of the model at each time the data for a patient is updated, with observations occurring 24 hours prior to a target event marked as 1 and otherwise marked as 0. Encounter level describes the model performance across the entire hospital encounter for one patient. It refers to the maximum model score during the patient's stay, occurring between admission and at least 30 minutes (or longer for different minimal lead times; see the section Comparison of PICTURE to EDI in a Non–COVID-19 Cohort) before the first event. The target in this case is 1 if the patient ever met an outcome condition during their stay, and 0 otherwise.

Comparison of PICTURE and EDI

Since the EDI makes a prediction every 15 minutes, we simulated how the PICTURE score, calculated at irregular intervals each time a new data point arrives, would align with the EDI. This limited the available number of encounters to 21,740 in the 2019 test set and 607 encounters in the COVID-19 cohort. The PICTURE scores were merged onto EDI values by taking the most recent PICTURE prediction before the EDI prediction. This was to give the EDI any advantages in the alignment procedure. Figure 2 displays a visual schematic of this alignment. We then evaluated the two models using the same observation-level and encounter-level methods described in the previous section.

Figure 2. Alignment of PICTURE predictions to EDI scores. Although the PICTURE system outputs predictions each time a new observation (eg, a new vital sign) is input in to the system, the EDI score is generated every 15 minutes. To give the EDI any potential advantage, PICTURE scores are aligned to EDI scores by selecting the most recent PICTURE score before each EDI prediction. In both cases, observations occurring 30 minutes before the target and after are excluded (red). For the patients who did not experience an adverse event, the maximum score was calculated across the entire encounter. EDI: Epic Deterioration Index; PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.



Performance Measures

AUROC and AUPRC were used as the primary criteria for comparison between the models. AUROC can be interpreted as the probability that two randomly chosen observations (one with a positive target, the other negative) are ranked in the correct order by the model prediction score. AUPRC describes the average positive predictive value (PPV) across the range of sensitivities. We also calculated 95% CIs for encounter-level statistics with a bootstrap method using 1000 replications to compute pivotal CIs. For observation-level statistics, block bootstrapping was used to ensure randomization between encounters and within the observations of an encounter. *P* values for AUROC differences were computed by counting the fraction of bootstrapped test statistics less than 0. If there were no simulations where the test statistic was greater than 0, the *P* value was recorded as $P<.001$.

Feature Ranking and Prediction Explanation

Despite the many benefits yielded by increasingly advanced machine learning models, use of these models in the medical field has lagged behind other fields. One contributing factor is their complexity, which make the resulting predictions difficult to interpret and in turn make it difficult to build clinician trust [20]. To better provide insight into the PICTURE predictions, tree-based Shapley values were calculated for each observation. Borrowed from game theory, Shapley values describe the relative contribution of a feature to the model's prediction [14,21]. Positive values denote features that influenced the model toward a high prediction score (here indicating a higher likelihood of an adverse event), while negative values indicate the feature pushed the model toward a lower prediction score. The sum of the Shapley values across a single prediction plus the mean log-odds probability of the model is proportional to the log-odds of the prediction probability. Shapley values can be used to provide insight into individual model predictions or aggregated to visualize global variable importance.

Calibration and Alert Thresholds

Neither PICTURE nor the EDI are calibrated scores—that is, even though their output ranges from 0 to 1 (or 0 to 100 in the case of EDI), these values do not reflect a probability of deterioration [11]. Furthermore, both PICTURE and the EDI were trained on cohorts of non–COVID-19 patients. which have a much lower event rate and therefore may require a different alert threshold. A calibration curve depicting PICTURE and EDI score quantiles against calculated risk is used to demonstrate the deviation of PICTURE and EDI scores from an estimated probability. Several simulated PICTURE alarm thresholds are then examined, calculated by aligning them to the EDI threshold suggested in Singh et al [11] via sensitivity, specificity, PPV, and negative predictive value (NPV). The performance at these thresholds simulates when and how often a clinician would receive alerts. Data from an example patient

is also highlighted to demonstrate how these alert thresholds and Shapley values may interact to provide actionable insights to clinicians.

Results

Validation of PICTURE Performance in a Non-COVID-19 Cohort

The ability of the PICTURE model to accurately predict the composite target was first assessed using the 33,472 encounters in the holdout test set from 2019. To provide a baseline for comparison, NEWS scores were calculated alongside each PICTURE prediction output. The observation-level and encounter-level AUROC and AUPRC are presented with 95% CIs in Table 2. The observation-level event rate can be interpreted as the fraction of individual observations during which an adverse event occurred within 24 hours, while the encounter-level event rate refers to the proportion of hospital encounters experiencing such an event. The difference in AUROC between PICTURE and NEWS was 0.068 (95% CI 0.058-0.078; *P*<.001) on the observation level and 0.064 (95% CI 0.055-0.073; *P*<.001) on the encounter level. The difference in AUPRC was similarly significant, at 0.041 (95% CI 0.031-0.050; *P*<.001) and 0.141 (95% CI 0.120-0.162; *P*<.001) on the observation and encounter levels, respectively.

Table 2. Evaluation of PICTURE (performance in a non–COVID-19 cohort).

Granularity and analytic	AUROC[a] (95% CI[b])	*P* value[c] (AUROC)	AUPRC[d] (95% CI)	*P* value (AUROC)	Event rate (%)
Observation		<.001		<.001	1.01
PICTURE[e]	0.821 (0.810-0.832)		0.099 (0.085-0.110)		
NEWS[f,g]	0.753 (0.741-0.765)		0.058 (0.049-0.064)		
Encounter (n=33,472)		<.001		<.001	3.99
PICTURE	0.846 (0.834-0.858)		0.326 (0.301-0.351)		
NEWS	0.782 (0.768-0.795)		0.185 (0.165-0.203)		

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

[b]95% CIs were calculated using a block bootstrap with 1000 replicates. In the case of the observation level, this bootstrap was blocked on the encounter level.

[c]*P* values are calculated using the bootstrap method outlined in the section Performance Measures.

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

[e]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[f]NEWS: National Early Warning Score.

[g]NEWS is used as a baseline for comparison.

Comparison of PICTURE to EDI in a Non–COVID-19 Cohort

PICTURE was then compared to the EDI model on non–COVID-19 patients in the same holdout test set from 2019. Due to limitations in available EDI scores, the number of encounters was restricted to 21,740. These time-matched scores were again evaluated using AUROC and AUPRC on the observation and encounter levels (Table 3). Panels A and B in Figure 3 display the associated receiver operating characteristic (ROC) and precision-recall (PR) curves for the observation-level performance. The difference in AUROC and AUPRC between PICTURE and the EDI reached significance on both the observation level (AUROC 0.056, 95% CI 0.044-0.068; *P*<.001; AUPRC 0.033, 95% CI 0.021-0.045; *P*<.001) and the encounter level (AUROC 0.056, 95% CI 0.046-0.065; *P*<.001; AUPRC 0.094, 95% CI 0.069-0.119; *P*<.001). NEWS results were similarly significant and are included in Table 3 for comparison.

Table 3. Comparison of PICTURE and the EDI in a non–COVID-19 cohort.

Granularity and analytic	AUROC[a] (95% CI)	*P* value (AUROC)[b]	AUPRC[c] (95% CI)	*P* value (AUPRC)	Event rate (%)
Observation					0.77
PICTURE[d]	0.819 (0.805-0.834)	• vs EDI[e]: <.001 • vs NEWS[f]: <.001	0.115 (0.096-0.130)	• vs EDI: <.001 • vs NEWS: <.001	
EDI	0.763 (0.746-0.781)	• vs NEWS: .01	0.081 (0.066-0.094)	• vs NEWS: <.001	
NEWS	0.745 (0.729-0.761)	• N/A[g]	0.062 (0.051-0.072)	• N/A	
Encounter (n=21,740)					4.21
PICTURE	0.859 (0.846-0.873)	• vs EDI: <.001 • vs NEWS: <.001	0.368 (0.335-0.400)	• vs EDI: <.001 • vs NEWS: <.001	
EDI	0.803 (0.788-0.821)	• vs NEWS: .15	0.274 (0.244-0.301)	• vs NEWS: <.001	
NEWS	0.797 (0.781-0.814)	• N/A	0.229 (0.204-0.254)	• N/A	

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

[b]*P* values reflect the difference in AUROC or AUPRC.

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

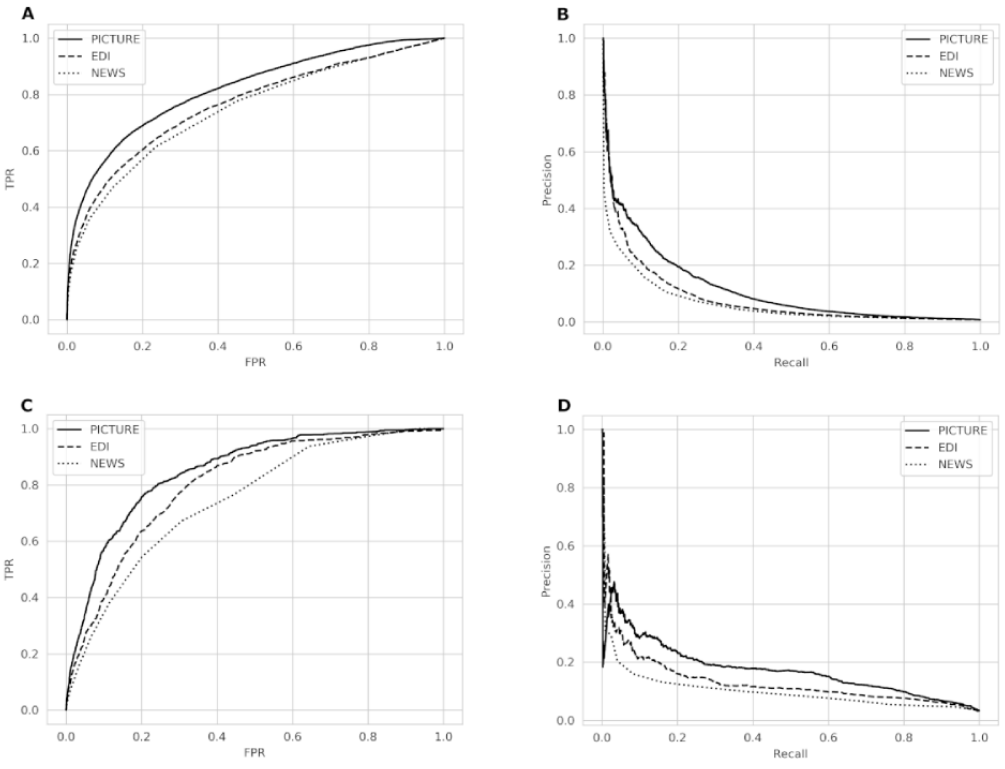
[d]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[e]EDI: Epic Deterioration Index.

[f]NEWS: National Early Warning Score.

[g]N/A: not applicable.

Figure 3. Comparison of PICTURE and the EDI. Panel A: receiver operating characteristic (ROC) curves for PICTURE, EDI, and NEWS models in the non–COVID-19 cohort. PICTURE area under the curve (AUC): 0.819; EDI AUC: 0.763; NEWS AUC: 0.745. Panel B: Precision-recall (PR) curves for the two models in the non–COVID-19 cohort. PICTURE AUC: 0.115; EDI AUC: 0.081; NEWS AUC: 0.062. Panel C: ROC curves for PICTURE, EDI, and NEWS models in the COVID-19 cohort. PICTURE AUC: 0.849; EDI AUC: 0.803; NEWS AUC: 0.746. Panel D: PR curves for the two models. PICTURE AUC: 0.173; EDI AUC: 0.131; NEWS AUC: 0.098 in the COVID-19 cohort. All curves represent observation-level analysis. EDI: Epic Deterioration Index; FPR: false-positive rate; NEWS: National Early Warning Score; PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events; TPR: true-positive rate.



In addition to classification performance, lead time represents another critical component of a predictive analytics' utility. Lead time refers to the amount of time between the alert and the actual event, and it determines how much time clinicians have to act on the model's recommendations. We assessed the model's relative performance at different lead times in a threshold-independent manner by excluding data occurring 0.5 hours, 1 hour, 2 hours, 6 hours, 12 hours, and 24 hours before an adverse event and calculating encounter-level performance (Table 4). In our cohort, PICTURE's AUROC and AUPRC were significantly higher (P<.001) than the EDI model even when considering predictions made 24 hours or more before the actual event.

Table 4. Lead time analysis in non–COVID-19 cohort.[a]

Lead time (hours)	AUROC[b] (95% CI)		AUPRC[c] (95% CI)		Event rate (%)	Sample size, n
	PICTURE[d]	EDI[e]	PICTURE	EDI		
0.5	0.859 (0.846-0.873)	0.803 (0.787-0.820)	0.368 (0.336-0.400)	0.274 (0.244-0.302)	4.21	21,636
1	0.850 (0.835-0.864)	0.795 (0.778-0.811)	0.346 (0.315-0.379)	0.254 (0.227-0.280)	4.18	21,636
2	0.838 (0.823-0.853)	0.784 (0.767-0.802)	0.321 (0.292-0.352)	0.238 (0.210-0.265)	4.14	21,622
6	0.825 (0.810-0.840)	0.768 (0.750-0.787)	0.280 (0.249-0.310)	0.210 (0.184-0.237)	3.92	21,572
12	0.817 (0.801-0.832)	0.767 (0.749-0.786)	0.247 (0.215-0.275)	0.183 (0.159-0.207)	3.67	21,515
24	0.808 (0.790-0.826)	0.759 (0.740-0.779)	0.205 (0.172-0.230)	0.144 (0.121-0.164)	3.24	21,419

[a]The performance of the two models (encounter level) at various lead times were assessed by evaluating the maximum prediction score prior to *x* hours before the given event, with *x* ranging in progressively greater intervals from 0.5 to 24. On this cohort of non–COVID-19 patients, PICTURE consistently outperformed the EDI. At each level of censoring, the *P* value when comparing PICTURE to the EDI was <.001.

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

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

[d]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[e]EDI: Epic Deterioration Index.

Comparison of PICTURE to EDI in Patients With COVID-19

When applied to patients testing positive for COVID-19, PICTURE performed similarly well. PICTURE scores were again aligned to EDI scores using the process outlined in the section Comparison of PICTURE and EDI. This resulted in the inclusion of 607 encounters. Table 5 presents AUROC and AUPRC values for PICTURE and the EDI on both the observation and encounter levels with 95% CIs and includes NEWS scores for comparison. Panels C and D in Figure 3 display the associated ROC and PR curves. The difference in AUROC and AUPRC between PICTURE and the EDI reached statistical significance (α=5%) on the observation level (AUROC 0.046, 95% CI 0.021-0.069; P<.001; AUPRC 0.043, 95% CI 0.006-0.071; P=.002) and the encounter level (AUROC 0.093, 95% CI 0.066-0.118; P<.001; AUPRC 0.155, 95% CI 0.089-0.204; P<.001). Of note, the EDI results at the observation level (AUROC 0.803, 95% CI 0.771-0.838) were similar to those described in a previous validation (AUROC 0.76, 95% CI 0.68-0.84), although with a smaller confidence interval due to a larger sample size [11]. The differences in AUROC and AUPRC between PICTURE and NEWS also reached significance (α=5%) in patients with COVID-19, both on the observation level (AUROC 0.104, 95% CI 0.075-0.129; P<.001; AUPRC 0.076, 95% CI 0.033-0.105; P<.001) and the encounter level (AUROC 0.122, 95% CI 0.090-0.154; P<.001; AUPRC 0.224, 95% CI 0.151-0.290; P<.001).

As with the non–COVID-19 cohort, a similar lead time analysis was then performed to assess the performance of PICTURE and EDI when making predictions further in advance. Thresholds were again set at 0.5 hours, 1 hour, 2 hours, 6 hours, 12 hours, and 24 hours before the event, and observations occurring after this cutoff were excluded. In our cohort, PICTURE again outperformed the EDI even when making predictions 24 hours in advance (Table 6).

Table 5. Comparison of PICTURE and the EDI in patients testing positive for COVID-19.

Granularity and analytic	AUROC[a] (95% CI)	*P* value (AUROC)	AUPRC[b] (95% CI)	*P* value (AUPRC)	Event rate (%)
Observation					3.20
PICTURE[c]	0.849 (0.820-0.878)	• vs EDI[d]: <.001 • vs NEWS[e]: <.001	0.173 (0.116-0.211)	• vs EDI: .002 • vs NEWS: <.001	
EDI	0.803 (0.772-0.838)	• vs NEWS: <.001	0.131 (0.087-0.163)	• vs NEWS: .002	
NEWS	0.746 (0.708-0.783)	• N/A[f]	0.098 (0.066-0.122)	• N/A	
Encounter (n=607)					20.6
PICTURE	0.895 (0.868-0.928)	• vs EDI: <.001 • vs NEWS: <.001	0.665 (0.590-0.743)	• vs EDI: <.001 • vs NEWS: <.001	
EDI	0.802 (0.762-0.848)	• vs NEWS: .05	0.510 (0.438-0.588)	• vs NEWS: .02	
NEWS	0.773 (0.732-0.818)	• N/A	0.441 (0.364-0.510)	• N/A	

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

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

[c]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[d]EDI: Epic Deterioration Index.

[e]NEWS: National Early Warning Score.

[f]N/A: not applicable.

Table 6. Lead time analysis in COVID-19 cohort.[a]

Lead time (hours)	AUROC[b] (95% CI)		AUPRC[c] (95% CI)		Event rate (%)	Sample size, n
	PICTURE[d]	EDI[e]	PICTURE	EDI		
0.5	0.895 (0.867-0.926)	0.802 (0.761-0.842)	0.665 (0.586-0.739)	0.510 (0.436-0.587)	20.6	607
1	0.887 (0.860-0.918)	0.793 (0.753-0.836)	0.631 (0.553-0.710)	0.491 (0.418-0.570)	20.5	606
2	0.870 (0.840-0.901)	0.794 (0.754-0.833)	0.598 (0.518-0.675)	0.478 (0.400-0.555)	20.1	603
6	0.847 (0.813-0.885)	0.769 (0.729-0.813)	0.552 (0.474-0.639)	0.435 (0.354-0.517)	19.3	597
12	0.821 (0.783-0.863)	0.752 (0.708-0.798)	0.497 (0.411-0.577)	0.403 (0.333-0.480)	17.9	587
24	0.808 (0.767-0.856)	0.740 (0.690-796)	0.443[f] (0.344-0.529)	0.370 (0.289-0.459)	16.0	574

[a]The performance of the two models (encounter level) at various lead times were again assessed by evaluating the maximum prediction score prior to x hours before the given event, with x ranging in progressively greater intervals from 0.5 to 24. On this cohort of non–COVID-19 patients, PICTURE consistently outperformed the EDI. At each level of censoring, the *P* value when comparing PICTURE to the EDI was <.001 unless otherwise marked.

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

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

[d]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[e]EDI: Epic Deterioration Index.

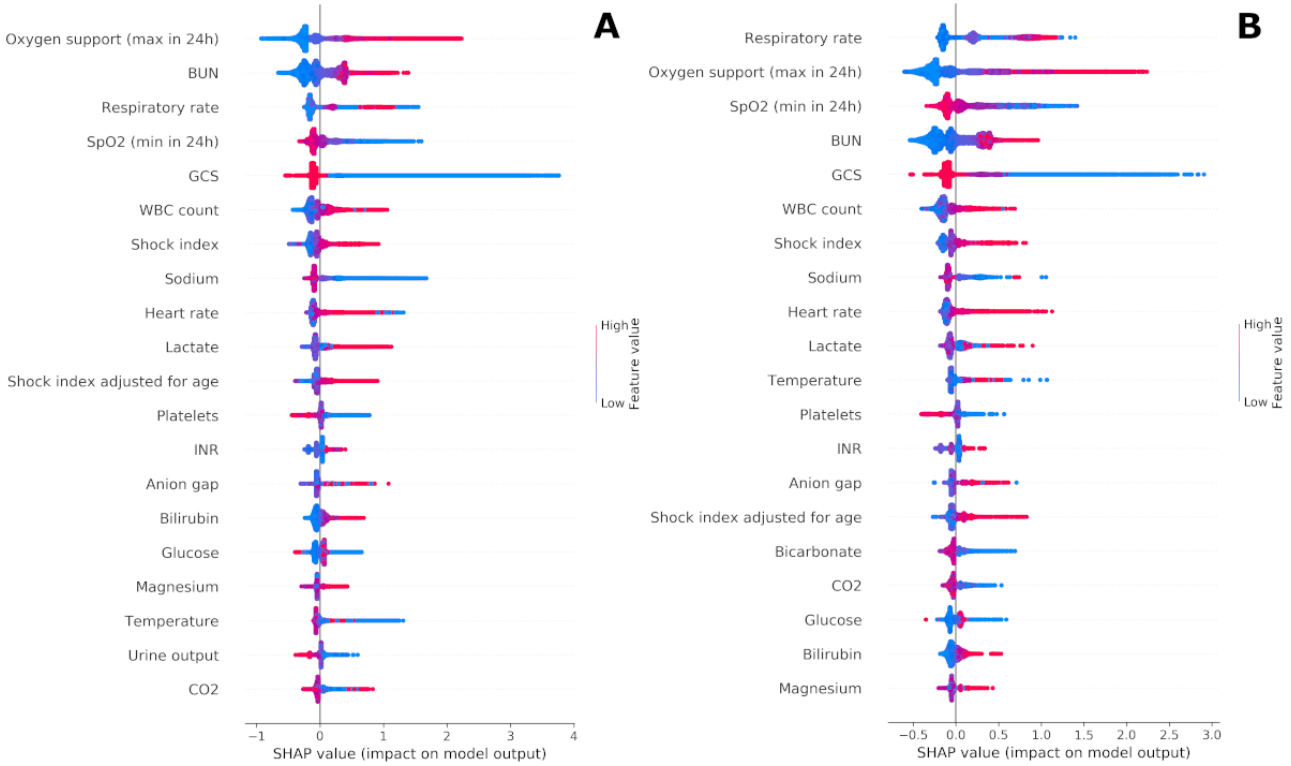
[f]*P*=.001.

Explanations of Predictions

To provide clinicians with a description of factors influencing a given PICTURE score, we used Shapley values computed at each observation. Figure 4 depicts an aggregated summary of the 20 most influential features in the 2019 test set (panel A) and in the COVID-19 set (panel B). Positive Shapley values indicate that the variable pushed the PICTURE score toward a positive decision (ie, predicting an adverse event). Although many of the feature rankings appear similar between the 2019 and COVID-19 cohorts, we noted that respiratory variables such as respiratory rate, oxygen support, and SpO_2 played a more pronounced role in predicting adverse events in COVID-19 positive patients than in non–COVID-19 patients. Multimedia Appendix 1 Figure S1 [22]. provides expanded detail on several of the variables (eg, respiratory rate and temperature) whose Shapley values do not appear to monotonically increase with their magnitude. One point of note is that the amount of oxygen support played a significant role in both cohorts. Although the EDI does not use the amount of oxygen support as a continuous

variable, it does have a feature termed “oxygen requirement” [11]. To demonstrate that the observed improvement of PICTURE over the EDI is not driven solely by this additional information, oxygen support was binarized and the PICTURE model retrained. Although performance did decrease, indicating that the inclusion of oxygen support as a continuous variable is useful in predicting deterioration, PICTURE still outperformed the EDI on both the non–COVID-19 (difference in AUROC 0.057, AUPRC 0.082) and COVID-19 (difference in AUROC 0.035, AUPRC 0.050) cohorts. Thus, oxygen support alone does not account for the difference between PICTURE and EDI performance.

Figure 4. Shapley summary plots. Panel A depicts an aggregated summary plot of the Shapley values from the 2019 test set, while panel B corresponds to COVID-19 positive patients. The 20 most influential features are ranked from top to bottom, and the distribution of Shapley values across all predictions are plotted. The magnitude of the Shapley value is displayed on the horizontal axis, while the value of the feature itself is represented by color. For example, a large amount of oxygen support over 24 hours (red) in panel A was associated with a highly positive influence on the model, while low to no oxygen support (blue) pushed the model back toward 0. BUN: blood urea nitrogen; GCS: Glasgow Coma Scale; INR: international normalized ratio; SHAP: Shapley; WBC: white blood cells.

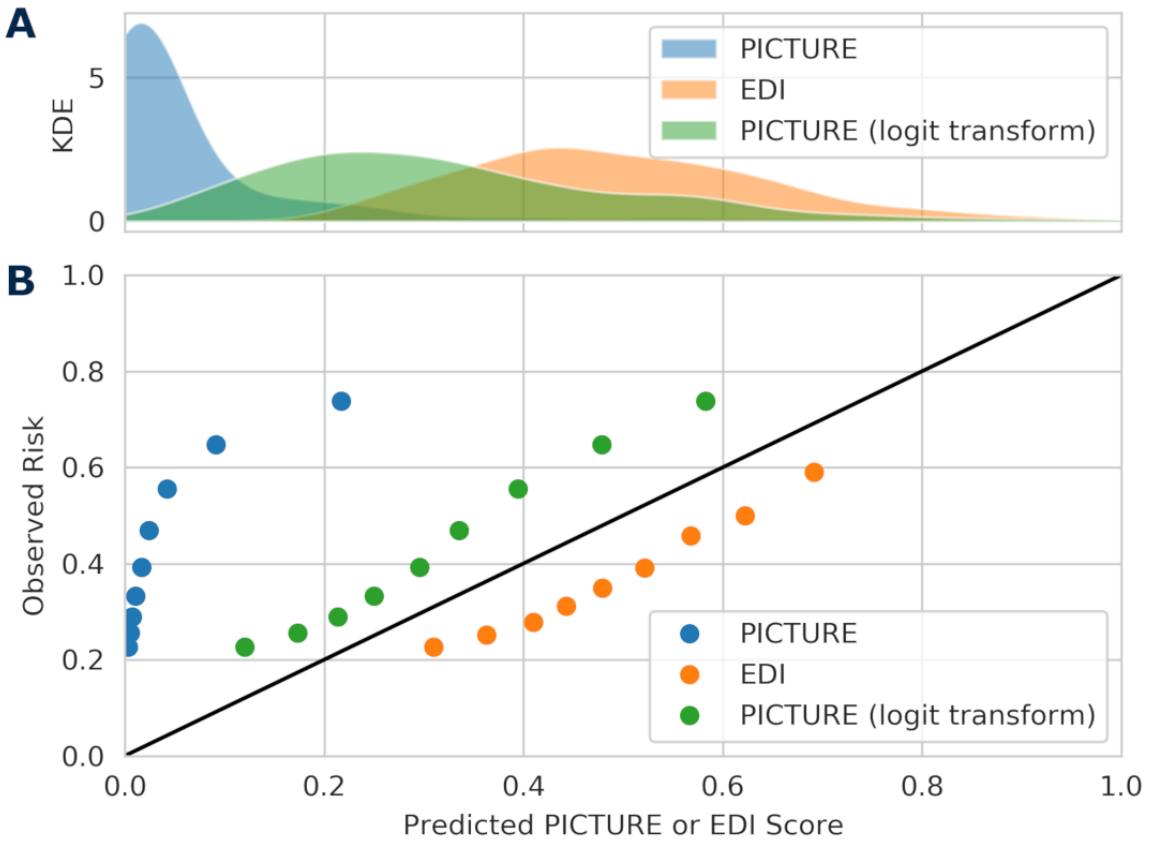


Calibration and Alert Thresholds

Both PICTURE and the EDI return scores indicate a patient’s risk of deterioration; however, neither score is calibrated as a probability. Therefore, alert thresholds may provide a convenient mechanism to decide whether or not to alert a clinician that their patient is at increased risk. A previous study assessing the use of the EDI in patients with COVID-19 found that an EDI score of 64.8 or greater to be an actionable threshold to identify patients at increased risk [11]. As PICTURE scores lie on a different scale than the EDI, calibration is required to simulate PICTURE alert thresholds.

Figure 5 depicts the distribution of PICTURE and EDI scores and a calibration curve comparing quantiles of PICTURE and EDI with observed risk. In this figure, EDI scores are rescaled from 0-100 to 0-1, while raw PICTURE scores are presented alongside a transformed score using a monotonically increasing function (logit transform) and scaled to the range 0-1. Based on this curve, the EDI appears to overestimate risk, while PICTURE may underestimate risk. However, neither metric is intended to reflect a probability. To more closely approximate a probability, techniques such as Platt scaling or isotonic regression may improve calibration in the future. Multimedia Appendix 1 Figure S2 illustrates the distribution of scores separated by positive and negative outcomes, and indicates that the PICTURE score may provide more separation between patients, something that the EDI has previously been demonstrated to struggle with [11].

Figure 5. Distribution of scores and calibration curve. Panel A presents a KDE of the distribution of PICTURE and EDI scores. In addition to raw PICTURE scores, logit-transformed scores are also included. Panel B depicts quantiles of PICTURE and EDI scores (0.1, 0.2, 0.3,...0.9) against observed risk. Neither PICTURE nor the EDI are calibrated as probabilities, and as such, the use of set alarm thresholds may be useful to help alert clinicians when their patient is at an increased risk. EDI: Epic Deterioration Index; KDE: kernel density estimate; PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.



To simulate when a clinician might receive an alert from the PICTURE system, four thresholds were selected, aligned based on the observed sensitivity, specificity, PPV, and NPV of the EDI score using the 64.8 value posed by Singh et al [11]. As an example, the *aligned by sensitivity* threshold listed in Table 7 was derived by determining the PICTURE threshold that had a sensitivity of 0.448, matching that of the EDI. Each of these thresholds, and their performances measured via F1 score, are compared to the EDI and are included in Table 7. The workup to detection ratio is calculated as 1 / PPV and indicates the number of false alerts a clinician might receive for each true positive [6]. For PICTURE, the workup to detection ratio ranged from 1.46 to 1.52 on the encounter level depending on the threshold used, compared to the EDI's 1.71. The median time between alert and adverse event according to each threshold is also displayed. Confusion matrices describing the performance of the model at each threshold are included in Multimedia Appendix 1 (Table S3).

Table 7. Alert thresholds and median lead time.[a]

Score	Threshold source	Threshold value	Sensitivity	Specificity	PPV[b]	NPV[c]	WDR[d]	F1 score[e]	Lead time[f] (h:min), median (IQR)
EDI[g]	Singh et al [11]	64.8	0.448	0.917	0.583	0.865	1.71	0.507	32:26 (4:37-66:08)
PICTURE[h]									
	Align by sensitivity	0.165	N/A[i]	0.946	0.683	0.869	1.46	0.541	40:14 (7:51-67:50)
	Align by specificity	0.097	0.616	N/A	0.658	0.902	1.52	0.636	40:04 (7:44-91:00)
	Align by PPV	0.048	0.792	0.851	N/A	0.940	N/A	0.668	54:10 (29:26-115:50)
	Align by NPV	0.173	0.432	0.946	0.675	N/A	1.48	0.527	41:40 (7:31-68:30)

[a]Sensitivity, specificity, PPV, and NPV were calculated for the EDI at a threshold of 64.8 as suggested in Singh et al [11] and based off encounter-level performance. PICTURE thresholds were then aligned to match these statistics. The WDR is also calculated as 1 / PPV and represents the number of false alarms received for each true positive. This value is important in limiting alert fatigue for clinicians and indicates that PICTURE may yield as much as 17% fewer false alarms for each true positive.

[b]PPV: positive predictive value.

[c]NPV: negative predicative value.

[d]WDR: workup to detection ratio.

[e]F1 scores were calculated as the harmonic mean between PPV and sensitivity.

[f]Lead times were determined using the intersection of true positives between PICTURE and the EDI, and were calculated as the time between a patient first crossing the threshold and their first deterioration event.

[g]EDI: Epic Deterioration Index.

[h]PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events.

[i]N/A: not applicable.

Discussion

Validation of PICTURE Performance in Non–COVID-19 Cohort

PICTURE makes a prediction at every observation for the features included. A natural starting point for the assessment of PICTURE's performance is at this level of granularity. Using the general structure outlined in Gillies et al [13], we updated the PICTURE model to reflect the target outcomes of death, ICU transfer or accommodation, and mechanical ventilation within 24 hours. This updated model was tested on data from 33,472 encounters in 2019 to ensure its performance (observation-level AUROC 0.821) was reasonably consistent with that described in Gillies et al [13]. It was also compared to the NEWS scores at simultaneous time points and was found to have significantly outperformed NEWS (AUROC 0.753). These results confirm the findings in Gillies et al [13] using 2019 data instead of 2018 data. They also provide a baseline of comparison as we move to predictions made at uniform intervals instead of every observation.

Comparison of PICTURE to EDI in a Non–COVID-19 Cohort

The EDI does not make predictions at every feature observation; instead, it makes predictions every 15 minutes. To provide a direct comparison to the EDI, we subset the PICTURE scores and time-matched them to the EDI scores as described in the section Performance Measures. PICTURE significantly outperformed the EDI on this cohort of non–COVID-19 patients, with an observation-level AUROC of 0.819 compared to the EDI's AUROC of 0.763. This performance gap extended out over multiple lead times, and even when restricted to data collected 24 hours or more before the adverse event, PICTURE's performance remained high with an AUROC of 0.808, while the EDI's AUROC dropped to 0.759. These results suggest that using PICTURE, instead of the EDI, at the University of Michigan hospital will lead to less false alarms. PICTURE maintained the performance improvement even as the models were forced to make predictions with longer times before the adverse event.

Comparison of PICTURE to EDI in Patients With COVID-19

As the EDI has increasingly been investigated as a feasible metric to gauge deterioration risk in patients with COVID-19 [11], we sought to apply our own deterioration model, PICTURE, to a cohort of patients with COVID-19. Although both models were trained and validated in non–COVID-19 general ward patients, their performance on our cohort of patients with COVID-19 was reasonably consistent with their respective results on our non–COVID-19 cohort. Even with a lower sample size (n=607 encounters), PICTURE significantly (P=.002) outperformed the EDI with an observation-level AUROC of 0.849 compared to the EDI's AUROC of 0.803. PICTURE's lead was again maintained 24 hours or more before the adverse event, with an AUROC of 0.808 versus the EDI's AUROC of 0.740. These results suggest that using PICTURE

instead of the EDI for patients with COVID-19 will lead to less alarm fatigue.

One important point of discussion is the considerably higher rate of deterioration observed in patients with COVID-19 (20.6% vs 4.21% of encounters). This is likely due to a combination of the severity of the virus when compared to a general ward population and the aggressive treatment regimen endorsed by clinicians facing a disease that, during the early phases of the pandemic, represented many unknowns. Therefore, the threshold selection presented in the section Calibration and Alert Thresholds may differ between COVID-19 and general ward patients. The performance of the PICTURE analytic (as measured by AUROC) increased slightly (though with overlapping 95% CIs) when applied to patients with COVID-19 versus the general test set, indicating that patients with COVID-19 may represent a slightly easier classification task. This is supported by the fact that the EDI also performed better on the COVID-19 cohort when measured by observation-level AUROC (0.763 vs 0.803), though this increase was not sustained in the encounter-level results (AUROC 0.803 vs 0.802).

Explanations of Predictions

One key feature of the PICTURE model is its use of Shapley values to help explain individual predictions to clinicians. These explanations help add interpretability to the model, allowing clinicians to evaluate individual model scores and identify potential next steps, follow-up tests, or treatment plans. Figure 4 depicts an aggregated summary of Shapley values across all observations in both the COVID-19 and non–COVID-19 cohorts. In non–COVID-19 patients, a high degree of oxygen support, high blood urea nitrogen (BUN), very high or very low respiratory rate, low SpO_2, and low GCS were the top five features most associated with high risk scores by the model. The COVID-19 cohort yielded the same top five features but reordered such that respiratory parameters (respiratory rate, oxygen support, and SpO_2) ranked above BUN and GCS. Of note, temperature was one of the few features that changed direction between the two cohorts. In non–COVID-19 patients, a high temperature was associated with low to moderate risk, whereas high temperatures in patients with COVID-19 tended to indicate those with the highest risk scores. The aggregate feature explanations are, in general, similar between the two cohorts and are largely consistent with clinician intuition. However, these few key differences may reflect some of the unique challenges faced when caring for patients with COVID-19.

Calibration and Alert Thresholds

Simulated alert thresholds were calculated based on the derived sensitivity, specificity, PPV, and NPV of the EDI threshold posited by Singh et al [11]. For each of the four thresholds, PICTURE outperformed the EDI according to the other four metrics as demonstrated in Table 7. For example, when the PICTURE alert threshold was adjusted such that its sensitivity matched the EDI's (0.448); the specificity (0.946), PPV (0.683), and NPV (0.869) were all higher than the EDI's (0.917, 0.583, and 0.865, respectively). Additionally, PICTURE's workup to detection ratio ranged from 1.46 to 1.52 on the encounter level depending on the threshold used, compared to the EDI's 1.71. This indicates that PICTURE may generate up to 17% fewer false positives for each true positive encounter.

Case Study Example

As a demonstration of the potential utility of PICTURE, an individual hospital encounter was selected, and the trajectories of PICTURE and the EDI are visualized in Figure 6. The EDI score threshold of 64.8, suggested by Singh et al [11], and the sensitivity-aligned and PPV-aligned PICTURE thresholds are also depicted. Note that the PICTURE score remains low until approximately 12.5 hours before the adverse event (in this case, transfer to an ICU level of care), where it crosses the PPV-aligned threshold. Approximately 11 hours before the event, the PICTURE score peaks at a value of 0.235 and exceeds the sensitivity-aligned threshold of 0.165. After the initial peak, the PICTURE score then remains elevated, staying above the PPV-aligned threshold of 0.048 until the patient is transferred. In contrast, the EDI score never exceeded its alert threshold, and it dropped when the PICTURE score increased.

To simulate what a clinician receiving an alert from PICTURE might encounter, the Shapley values explaining the PICTURE predictions at both alert thresholds are recorded in Table 8. Note that these explanations are dominated by respiratory features, though heart rate and temperature are also present. Although these features may seem obvious in predicting the need for ICU care, it is worth highlighting that the EDI did not identify this patient as being at risk.

Figure 6. Sample trajectory of one patient. Panel A depicts the PICTURE predictions over 27 hours before the patient is eventually transferred to an ICU level of care (green bar). Two possible alert thresholds are noted: one (red: 0.165) based on the EDI's sensitivity at a threshold of 64.8 (as suggested by Singh et al [11]), while the other (yellow: 0.048) is based on the EDI's PPV at this threshold. Note that PICTURE peaks above the sensitivity-based threshold approximately 11 hours in advance of the ICU transfer and then remains elevated over the PPV threshold until the transfer occurs. * and † represent the first time points that PICTURE crossed each threshold, referenced in Table 7. Panel B demonstrates the EDI over the same time range, with the threshold of 64.8 suggested by Singh et al [11]. The EDI did not identify this patient as being at risk. EDI: Epic Deterioration Index; ICU: intensive care unit; PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events; PPV: positive predictive value.

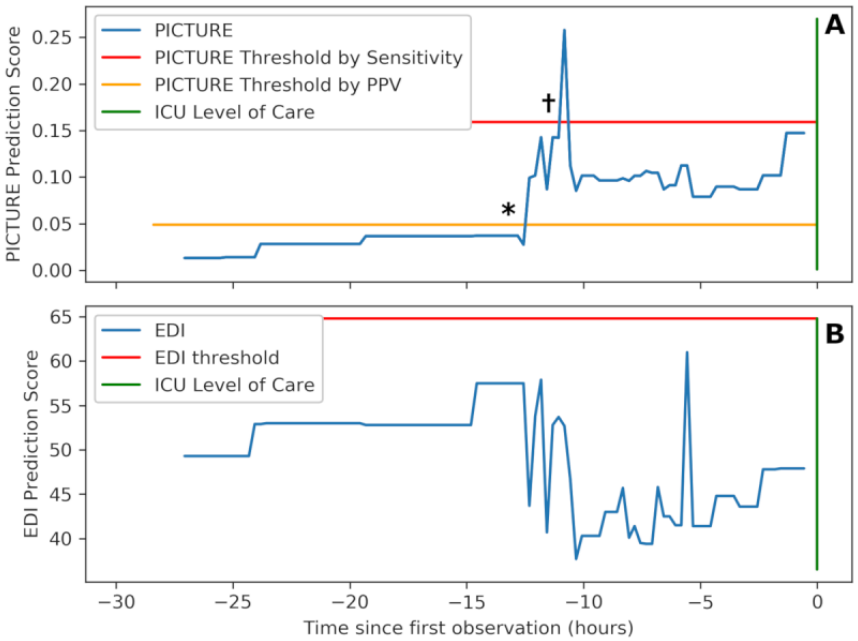


Table 8. Sample Predicting Intensive Care Transfers and Other Unforeseen Events explanations.

Rank and feature name[a]	Value	Median (IQR)[b]	Shapley score
Shapley values after PPV [c] threshold (t – 12.75 h)			
1. Oxygen supplementation (rolling 24 h max)	7 L/min	2.0 (0.0-3.0)	1.06
2. SpO_2[d] (rolling 24 h min)	85%	92.0 (90.0-94.0)	0.93
3. Respiratory rate	26 bpm	20.0 (18.0-20.0)	0.76
4. Temperature	39.1 ˚C	36.9 (36.8-37.2)	0.32
5. Protein level	5.7	6.0 (5.6-6.4)	0.13
Shapley values after sensitivity threshold (t – 11 h)			
1. Oxygen supplementation (rolling 24 h max)	35 L/min	2.0 (0.0-3.0)	1.93
2. SpO_2 (rolling 24 h min)	85%	92.0 (90.0-94.0)	1.09
3. Respiratory rate	24 bpm	20.0 (18.0-20.0)	0.73
4. Heart rate[e]	124 bpm	83.0 (74.0-92.0)	0.71
5. Temperature	39.1˚C	36.9 (36.8-37.2)	0.32

[a]The top 5 features corresponding to Predicting Intensive Care Transfers and Other Unforeseen Events predictions as it crosses the PPV-aligned threshold and the sensitivity-aligned threshold as noted in Figure 6. These predictions represent two possible locations where a clinician could receive an alert that their patient is deteriorating. Such information could be shared alongside the prediction score to provide better clinical utility to health care providers. Note that oxygenation (supplemental oxygen, SpO2, and respiratory rate) and temperature play a dominant role in both cases.

[b]The median and IQR are included for comparison, and are calculated using the COVID-19 data set.

[c]PPV: positive predictive value.

[d]SpO_2: oxygen saturation as measured by pulse oximetry.

[e]Heart rate represented the primary difference between these two time points. When the Predicting Intensive Care Transfers and Other Unforeseen Events score first exceeded the PPV threshold 12.5 hours before the intensive care unit transfer, the heart rate remained at 65 bpm and was not among the top features as measured by Shapley. At 11 hours before the event, when the Predicting Intensive Care Transfers and Other Unforeseen Events score was at its highest, the heart rate had jumped to 124 bpm and was the fourth-most influential feature as measured by Shapley values.

Limitations

This analysis is limited to a single academic medical center, and its generalizability to other health care systems will require future study. Our sample of patients with COVID-19 was also limited in size, limiting our power to detect differences between PICTURE and the EDI. Lastly, the thresholds presented in the section Calibration and Alert Thresholds may be different from those used in the general population due to the increased event rate. The thresholds may also require future tuning to suit the needs of individual units.

Conclusion

The PICTURE early warning system accurately predicts adverse patient outcomes including ICU transfer, mechanical ventilation, and death at Michigan Medicine. The ability to consistently anticipate these events may be especially valuable when considering a potential impending second wave of COVID-19 infections. The EDI is a widespread deterioration model, which has recently been assessed in a COVID-19 population. Both PICTURE and the EDI were trained using approximately 130,000 non–COVID-19 encounters for general deterioration and thus are not overfit to the COVID-19 population [11,12]. Using a head-to-head comparison, we demonstrated that PICTURE has higher performance than the EDI at a statistically significant level (α=5%) for both COVID-19 positive and non–COVID-19 patients. In addition, PICTURE was capable of accurately predicting adverse events as far as 24 hours before the event occurred. Lastly, PICTURE has the ability to explain individual predictions to clinicians by displaying those variables that most influenced its prediction using Shapley values.

Acknowledgments

This study was supported in part by the Michigan Institute for Data Science “Propelling Original Data Science (PODS) Mini-Grants for COVID-19 Research” award. AJA has received funding from NIH/NHLBI (F32HL149337).

Conflicts of Interest

CEG, RPM Jr, and KRW have submitted a patent regarding our machine learning methodologies presented in this paper through the University of Michigan’s Office of Technology Transfer.

Multimedia Appendix 1
Supplementary material.
[DOCX File , 24849 KB - medinform_v9i4e25066_app1.docx]

References

1. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
2. Smith MEB, Chiovaro JC, O'Neil M, Kansagara D, Quiñones AR, Freeman M, et al. Early warning system scores for clinical deterioration in hospitalized patients: a systematic review. Ann Am Thorac Soc 2014 Nov;11(9):1454-1465. [doi: 10.1513/AnnalsATS.201403-102OC] [Medline: 25296111]
3. Le Lagadec MD, Dwyer T. Scoping review: the use of early warning systems for the identification of in-hospital patients at risk of deterioration. Aust Crit Care 2017 Jul;30(4):211-218. [doi: 10.1016/j.aucc.2016.10.003] [Medline: 27863876]
4. McGaughey J, Alderdice F, Fowler R, Kapila A, Mayhew A, Moutray M. Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. Cochrane Database Syst Rev 2007 Jul 18(3):CD005529. [doi: 10.1002/14651858.CD005529.pub2] [Medline: 17636805]
5. Adverse events in hospitals: national incidence among Medicare beneficiaries - 10-year update. Office of Inspector General. 2018 Dec 17. URL: https://oig.hhs.gov/reports-and-publications/workplan/summary/wp-summary-0000328.asp [accessed 2020-06-17]
6. Linnen D, Escobar GJ, Hu X, Scruth E, Liu V, Stephens C. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. J Hosp Med 2019 Mar;14(3):161-169 [FREE Full text] [doi: 10.12788/jhm.3151] [Medline: 30811322]
7. Bapoje S, Gaudiani J, Narayanan V, Albert R. Unplanned transfers to a medical intensive care unit: causes and relationship to preventable errors in care. J Hosp Med 2011 Feb;6(2):68-72. [doi: 10.1002/jhm.812] [Medline: 21290577]
8. Young M, Gooder V, McBride K, James B, Fisher E. Inpatient transfers to the intensive care unit: delays are associated with increased mortality and morbidity. J Gen Intern Med 2003 Feb;18(2):77-83 [FREE Full text] [doi: 10.1046/j.1525-1497.2003.20441.x] [Medline: 12542581]
9. Rothman MJ. The emperor has no clothes. Crit Care Med 2019 Jan;47(1):129-130. [doi: 10.1097/CCM.0000000000003505] [Medline: 30557245]
10. Artificial intelligence triggers fast, lifesaving care for COVID-19 patients. Epic. URL: https://www.epic.com/epic/post/artificial-intelligence-epic-triggers-fast-lifesaving-care-covid-19-patients [accessed 2020-06-18]

11. Singh K, Valley TS, Tang S, Li BY, Kamran F, Sjoding MW, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized COVID-19 patients. Ann Am Thorac Soc 2020 Dec 24:1. [doi: 10.1513/AnnalsATS.202006-698OC] [Medline: 33357088]
12. Strickland E. AI may help hospitals decide which COVID-19 patients live or die. IEEE Spectrum. 2020 Apr 17. URL: https://spectrum.ieee.org/the-human-os/artificial-intelligence/medical-ai/ai-can-help-hospitals-triage-covid19-patients [accessed 2020-06-19]
13. Gillies C, Taylor DF, Cummings BC, Ansari S, Islim F, Kronick SL, et al. Demonstrating the consequences of learning missingness patterns in early warning systems for preventative health care: a novel simulation and solution. J Biomed Inform 2020 Oct;110:103528. [doi: 10.1016/j.jbi.2020.103528] [Medline: 32795506]
14. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020 Jan;2(1):56-67 [FREE Full text] [doi: 10.1038/s42256-019-0138-9] [Medline: 32607472]
15. Churpek MM, Yuen TC, Winslow C, Robicsek AA, Meltzer DO, Gibbons RD, et al. Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med 2014 Sep 15;190(6):649-655 [FREE Full text] [doi: 10.1164/rccm.201406-1022OC] [Medline: 25089847]
16. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Presented at: KDD '16; August 2016; San Francisco, CA p. 785-794. [doi: 10.1145/2939672.2939785]
17. Smith GB, Prytherch DR, Meredith P, Schmidt PE, Featherstone PI. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. Resuscitation 2013 Apr;84(4):465-470. [doi: 10.1016/j.resuscitation.2012.12.016] [Medline: 23295778]
18. National Early Warning Score (NEWS): standardising the assessment of acute-illness severity in the NHS. Royal College of Physicians. 2012. URL: https://www.rcplondon.ac.uk/file/32/download+&cd=4&hl=en&ct=clnk&gl=ca [accessed 2020-06-18]
19. Pimentel M, Redfern OC, Gerry S, Collins GS, Malycha J, Prytherch D, et al. A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: a multi-centre database study. Resuscitation 2019 Jan;134:147-156 [FREE Full text] [doi: 10.1016/j.resuscitation.2018.09.026] [Medline: 30287355]
20. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019 Jul 29;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
21. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017 Presented at: NIPS 2017; December 4-9, 2017; Long Beach, CA p. 4765-4774.
22. Ball JW, Dains JE, Flynn JA, Solomon BS, Steward RW. Seidel's Guide to Physical Examination, 9th edition. London: Elsevier Health Sciences; 2019.

Abbreviations

AUPRC: area under the precision-recall curve
AUROC: area under the receiver operating characteristic curve
BUN: blood urea nitrogen
EDI: Epic Deterioration Index
EHR: electronic health record
GCS: Glasgow Coma Scale
ICU: intensive care unit
NEWS: National Early Warning Score
NPV: negative predictive value
PICTURE: Predicting Intensive Care Transfers and Other Unforeseen Events
PODS: Propelling Original Data Science
PPV: positive predictive value
PR: precision-recall
ROC: receiver operating characteristic
SpO2: oxygen saturation as measured by pulse oximetry

Edited by C Lovis; submitted 16.10.20; peer-reviewed by S Shams, K Cato; comments to author 27.12.20; revised version received 15.01.21; accepted 03.04.21; published 21.04.21.

Please cite as:
Cummings BC, Ansari S, Motyka JR, Wang G, Medlin Jr RP, Kronick SL, Singh K, Park PK, Napolitano LM, Dickson RP, Mathis MR, Sjoding MW, Admon AJ, Blank R, McSparron JI, Ward KR, Gillies CE
Predicting Intensive Care Transfers and Other Unforeseen Events: Analytic Model Validation Study and Comparison to Existing Methods
JMIR Med Inform 2021;9(4):e25066
URL: https://medinform.jmir.org/2021/4/e25066
doi:10.2196/25066
PMID:33818393

Review

Machine Learning Models for Image-Based Diagnosis and Prognosis of COVID-19: Systematic Review

Mahdieh Montazeri[1], MSc; Roxana ZahediNasab[2], MSc; Ali Farahani[2], MSc; Hadis Mohseni[2], PhD; Fahimeh Ghasemian[2], PhD

[1]Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran

[2]Computer Engineering Department, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

Corresponding Author:
Fahimeh Ghasemian, PhD
Computer Engineering Department, Faculty of Engineering
Shahid Bahonar University of Kerman
Pajoohesh Sq, PO Box: 76169-14111
Kerman
Iran
Phone: 98 9133924837
Email: ghasemianfahime@uk.ac.ir

Abstract

Background: Accurate and timely diagnosis and effective prognosis of the disease is important to provide the best possible care for patients with COVID-19 and reduce the burden on the health care system. Machine learning methods can play a vital role in the diagnosis of COVID-19 by processing chest x-ray images.

Objective: The aim of this study is to summarize information on the use of intelligent models for the diagnosis and prognosis of COVID-19 to help with early and timely diagnosis, minimize prolonged diagnosis, and improve overall health care.

Methods: A systematic search of databases, including PubMed, Web of Science, IEEE, ProQuest, Scopus, bioRxiv, and medRxiv, was performed for COVID-19–related studies published up to May 24, 2020. This study was performed in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines. All original research articles describing the application of image processing for the prediction and diagnosis of COVID-19 were considered in the analysis. Two reviewers independently assessed the published papers to determine eligibility for inclusion in the analysis. Risk of bias was evaluated using the Prediction Model Risk of Bias Assessment Tool.

Results: Of the 629 articles retrieved, 44 articles were included. We identified 4 prognosis models for calculating prediction of disease severity and estimation of confinement time for individual patients, and 40 diagnostic models for detecting COVID-19 from normal or other pneumonias. Most included studies used deep learning methods based on convolutional neural networks, which have been widely used as a classification algorithm. The most frequently reported predictors of prognosis in patients with COVID-19 included age, computed tomography data, gender, comorbidities, symptoms, and laboratory findings. Deep convolutional neural networks obtained better results compared with non–neural network–based methods. Moreover, all of the models were found to be at high risk of bias due to the lack of information about the study population, intended groups, and inappropriate reporting.

Conclusions: Machine learning models used for the diagnosis and prognosis of COVID-19 showed excellent discriminative performance. However, these models were at high risk of bias, because of various reasons such as inadequate information about study participants, randomization process, and the lack of external validation, which may have resulted in the optimistic reporting of these models. Hence, our findings do not recommend any of the current models to be used in practice for the diagnosis and prognosis of COVID-19.

(JMIR Med Inform 2021;9(4):e25181) doi:10.2196/25181

KEYWORDS

machine learning; diagnosis; prognosis; COVID-19

Introduction

Since the COVID-19 outbreak was first reported in December 2019 in Wuhan, China, the number of people infected worldwide has exceeded 33 million (as of September 28, 2020) [1]. The World Health Organization declared COVID-19 as a global health emergency that requires international cooperation [2,3]. Despite many efforts to control the spread of the disease, many countries are facing a crisis of intensive care [4,5]. In order to reduce the burden on the health care system and provide the best possible care for patients, accurate and timely diagnosis and effective prognosis of COVID-19 is important and necessary. Moreover, early diagnosis of the disease helps health care providers prevent delays in providing the best possible treatment.

The diagnostic method currently used for COVID-19 is a positive result of a nucleic acid test such as real-time reverse transcription–polymerase chain reaction (RT-PCR) or next-generation sequencing [6]. Despite the advantages of this method, the number of false-negative test results due to unstable specimen processing is relatively high in clinical practice, which makes COVID-19 diagnosis difficult [7,8]. Moreover, laboratory testing for COVID-19 requires a rigorous platform, which is not assembled in all hospitals. Thus, COVID-19 testing may involve transfer of clinical specimens, which may delay diagnosis for days. Computed tomography (CT) plays a fundamental role in the diagnosis of disease progression, because of its excellent diagnostic accuracy and clinical outcomes [9]. For instance, lung CT images can be used to detect characteristic abnormalities associated with COVID-19 [10,11]. Characteristic imaging manifestations of COVID-19, such as ground-glass opacities, bilateral involvement, and peripheral distribution, have been described in various studies [12,13]. Consolidation, cavitation, and interlobular septal thickening imaging features have also been reported in some patients with COVID-19 [14,15].

Machine learning techniques have achieved considerable success in the field of medical imaging and image analysis owing to the use of deep learning technologies that allow for improved feature extraction [16,17]. Machine learning is a popular method of data analytics that uses different learning algorithms to teach computers to learn from data for performing related tasks. It is principally based on the learning methods and can be divided into three groups, namely, supervised (classification, regression, and ensembling), unsupervised (**association, clustering,** and **dimensionality reduction),** and reinforcement learning, with each category consisting of various methods for specific aims, such as instance-based algorithm, regression analysis, regularization, and classifiers for particular aims. Numerous studies have suggested the use of machine learning techniques in the diagnosis of diseases. For example, some studies have used deep learning techniques to diagnose and differentiate between bacterial and viral pneumonia using pediatric chest radiographic images [18,19]. Considerable effort has also been invested in diagnosing various chest CT imaging features that are characteristic of different diseases [20,21]. Various models ranging from rule-based systems to advanced machine learning models (deep learning) have been published in the context of the diagnosis and prognosis of COVID-19, which have substantially contributed to the field of health care by aiding the diagnosis and treatment of this disease and helped saved lives [22].

The objective of this systematic review was to identify publications in the existing literature that have used image processing methods based on CT images for the diagnosis and prognosis of COVID-19. We believe that this review would aid clinical practice by informing future research and development about improved diagnostic and treatment techniques for patients with COVID-19.

Methods

Information Source and Search Strategy

We conducted a systematic search of the databases, including PubMed, Web of Science, IEEE, ProQuest, Scopus, bioRxiv, and medRxiv, for articles published up to May 24, 2020. The study was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) guidelines [23]. We used two groups of keywords for searching these databases—keywords related to the novel coronavirus and those related to machine learning and image processing.

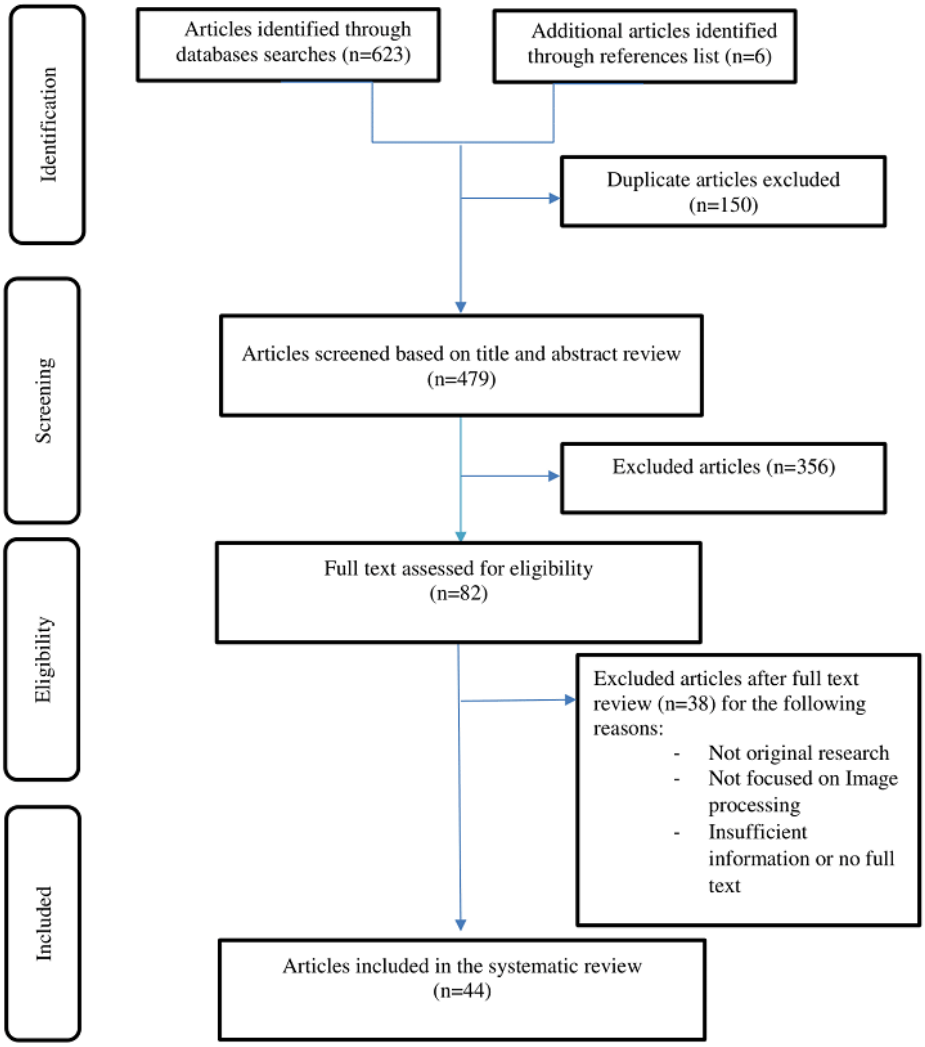
Inclusion and Exclusion Criteria

All studies that applied image processing techniques for the prediction and diagnosis of COVID-19 were considered. We included original research articles regardless of the language of publication. We excluded editorials, commentaries, letters, books, presentations, conference papers, and papers without full text or those with insufficient information. To prevent duplication in data collection, we also excluded all types of review articles.

Study Selection

The selection process was initiated by removing duplicated articles. Thereafter, two reviewers worked independently to screen the titles and abstracts of the selected articles against the eligibility criteria. We further excluded articles that did not apply image processing for the prediction and diagnosis of COVID-19. The detailed process regarding the selection of articles is presented in Figure 1. After the initial screening, the same authors independently reviewed the full text of the relevant articles. Any disagreements were resolved through mutual discussion. During the screening of the articles, the reviewers documented the reasons for the exclusion of each article. We used a free web and mobile application platform (Rayyan, Qatar Computing Research Institute) for the screening of articles [24].

Figure 1. Study identification and selection process.



Data Extraction and Synthesis

A standard data extraction form based on the Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies (CHARMS) checklist was used by five reviewers [25]. A data extraction form was used to extract specific details about each article. This form consisted of information on imaging modality, database, scope, setting, data source and outcome, sample size (including training, validation, and testing), machine learning technique, performance, validation type, risk of bias (Multimedia Appendix 1). We investigated several forms of validation, for example, external (ie, evaluation in an independent database) and internal validation (ie, bootstrap validation, cross validation, random training test splits, and temporal splits).

Risk of Bias Assessment

The risk of bias was assessed using the Prediction Model Risk of Bias Assessment Tool (PROBAST) [26].

Results

Overview

We retrieved 623 relevant studies through database searches. Six studies were identified from the reference lists of the selected publications. After title and abstract screening, 82 articles were selected for full-text assessment, which led to the exclusion of 38 articles due to various reasons.

In total, 44 studies were included in this systematic review (Figure 1). All included studies documented that patients' CT and chest x-ray (CXR) images were processed for segmentation and classification tasks to enable the diagnosis and prognosis of COVID-19. These studies described a total of 89 deep learning and machine learning models applied for COVID-19 screening of CT and CXR images (Table 1).

Table 1. Deep learning architecture and parameters.

Study	Network architecture	Optimizer	Learning rate	Batch size
[27]	U-Net	__[a]	—	—
[28]	Efficient Net B4+2 FC [29]	SGD[b]	1e-4	64
[30]	ResNet-50-2D [31]	—	—	—
[32]	CPM[c]-Nets [33]	—	—	—
[34]	U-Net (segmentation)	—	1e-5	32
[34]	ResNet 152 (classification)	—	1e-5	32
[35]	U-Net	—	—	—
[36]	AlexNet, GoogLeNet, and ResNet-18 + GAN[d]	SGD	0.01	64
[37]	AlexNet, VGG-16, VGG-19, SqueezeNet, GoogLeNet, MobileNet-V2, ResNet-18, ResNet-50, ResNet-101, and Xception	SGD	0.01	—
[38]	50×5 layers + 8FC[e] + 1 global average pooling + softmax 5 layers = (2 Conv + 3MP)	Adam	Optimize beside L2 regularization and momentum	32
[39]	VGG-19	Adam	0.001	15
[40]	DenseNet-201 + Inception_resnet_V2 + Inception_V3 + Mobilenet_V2 + ResNet-50 + VGG16 + VGG19 +	Adam	1e-5	32
[41]	2D (U-net + DRUNET + FCN[f] + SegNet + DeepLabv3)	SGD	0.01	4
[41]	3D (ResNet-18)	Adam	0.001	8
[42]	CNN[g] network base on the modification of ResNet-50 architecture	Rmsprop	1e-5	4
[43]	DenseNet like structure [44]	—	—	—
[45]	Model A, 22 layers	Adam	0.001	—
[45]	Model B, 28 layers	Adam	0.001	—
[45]	Model C, 29 layers	Adam	0.001	—
[46]	TB detection DL[h] model	—	—	—
[47]	MobileNetV2, SqueezeNet	SGD	1e-5	64
[48]	Darknet-19	Adam	3e-3	—
[49]	2D (ResNet-50)	—	—	—
[49]	3D (U-Net)	—	—	—
[50]	ResNet-18	Adam	0.001	16
[51]	MobileNetV2	—	—	—
[52]	DenseNet	SGD	—	32
[53]	GAN + VGG16	Adam	0.001	16
[54]	U-Net	Adam	1e-4	—
[55]	FC-DenseNet-103	Adam	1e-4	2
[55]	ResNet-18	Adam	1e-5	16
[56]	DeCoVNet	Adam	1e-5	32
[57]	3D-ResNet (prediction)	Momentum	1e-4	—
[57]	3D-UNet (segmentation)	Momentum	1e-4	—
[58]	ConvNet [59]	Adam	1e-4	64
[60]	INF-Net	Adam	1e-4	16
[60]	FCN8s	SGD	1e-10	16
[61]	UNet++ [62]	—	—	—

Study	Network architecture	Optimizer	Learning rate	Batch size
[63]	FCN-8s, U-Net, V-Net, and 3D U-Net++	Adam	1e-4	—
[64]	VB-Net	—	—	—
[65]	VB-Net	—	—	—
[66]	M-Inception (6Conv + 3MP + inception + softmax + 2FC)	—	—	—
[67]	VNET_IR_RPN [68]	—	—	—
[69]	DRE-NET (ResNet-50 as the backbone)	—	—	—
[70]	U-Net as segmentation	Adam	1e-5	1
[70]	DeconvNet as prediction	Adam	1e-5	1
[71]	MLP[i] + LSTM[j] (single layer) + FC + softmax	—	—	—
[72]	U-Net	—	—	—

[a]Not available.

[b]SGD: stochastic gradient descent.

[c]CPM: cross partial multiview networks.

[d]GAN: generative adversarial network.

[e]FC: fully connected layer.

[f]FCN: fully convolutional network.

[g]CNN: convolutional neural network.

[h]DL: deep learning.

[i]MLP: multilayer perceptron.

[j]LSTM: long short-term memory.

Dataset

Distribution of the 44 collected datasets showed that 12 (27%) studies used data on patients with COVID-19 from China; 3 (7%) studies used data on patients from China and USA [27,28,30]; 1 (2%) study used data on patients from China and Japan [32]; 1 (2%) study used data from China, USA, and Switzerland [34]; and 1 (2%) study used data from Italy [73], the Netherlands [35], and Canada [36]. Moreover, 11 (25%) studies were based on international data. Finally, the datasets used in 25 (56%) studies are publicly available, whereas those used in the rest of the studies (19/44, 43%) are nonpublic. The duration of follow-up was unclear for most studies. Only 2 (4%) studies reported follow-up time; the first one reported a follow-up of more than 5 days [28] and the other reported a follow-up of 3-6 days [37].

We categorized the reviewed studies (N=44) into three broad categories: (1) the *CT scan* category comprised 28 (63%) studies in which the models used chest CT images for abnormality analysis and COVID-19 diagnosis; (2) the *x-ray* category consisted of 14 (32%) studies in which the models use patients' CXR images; and (3) the *hybrid* category consisted of 3 (7%) studies in which the models use a combination of CT, CXR, lung ultrasound, and other information such as the patient's age and medical history.

Machine Learning Methods

Several machine learning techniques have been used for COVID-19 detection, prediction, and diagnosis. For the classification algorithms, the dataset is divided into training and test datasets. The model was developed using the training dataset, following which the validation of the training model was accomplished using the test dataset. For the segmentation algorithm, most studies used deep learning methods based on convolutional neural networks (CNNs) that have been used widely as a classification algorithm. In all, 40 studies used diagnostic models, whereas 4 studies used prognostic models for patients who had received a COVID-19 diagnosis [41,43,71,72]. Table 1 illustrates the deep learning architectures and hyperparameters used in the included studies using deep learning methods. In this table, the three most important parameters such as optimizer method, learning rate, and mini-batch size were considered. In the case of the optimizing algorithm Adam and RMSProp, all reported learning rates are initial values except in one study [29] that used a constant learning rate value.

Diagnostic Models to Detect COVID-19 in Patients With Suspected Infection

For better categorization among the various machine learning methods used in the studies analyzed, we classified the models into two groups: CNN-based models (n=31) and other machine learning algorithms (n=8). Among these, 31 studies used 61 CNN-based algorithms, which were further subdivided as follows: U-Net (n=10), ResNet (n=11), SqueezeNet (n=3), MobileNet (n=4), multiple types of VGG networks (n=4), GoogLeNet (n=2), and others (n=4). A total of 8 studies used 26 other machine learning methods, of which support vector machine (SVM) was the most commonly used algorithm as a classifier (n=5) [32,73-76], followed by random forest (n=1) [65,76], logistic regression (n=1) [34], and other machine learning algorithms (n=3). In addition, 1 study [77] used a multi-objective, differential, evolution-based algorithm to automatically build CNN. In addition, 4 models were developed

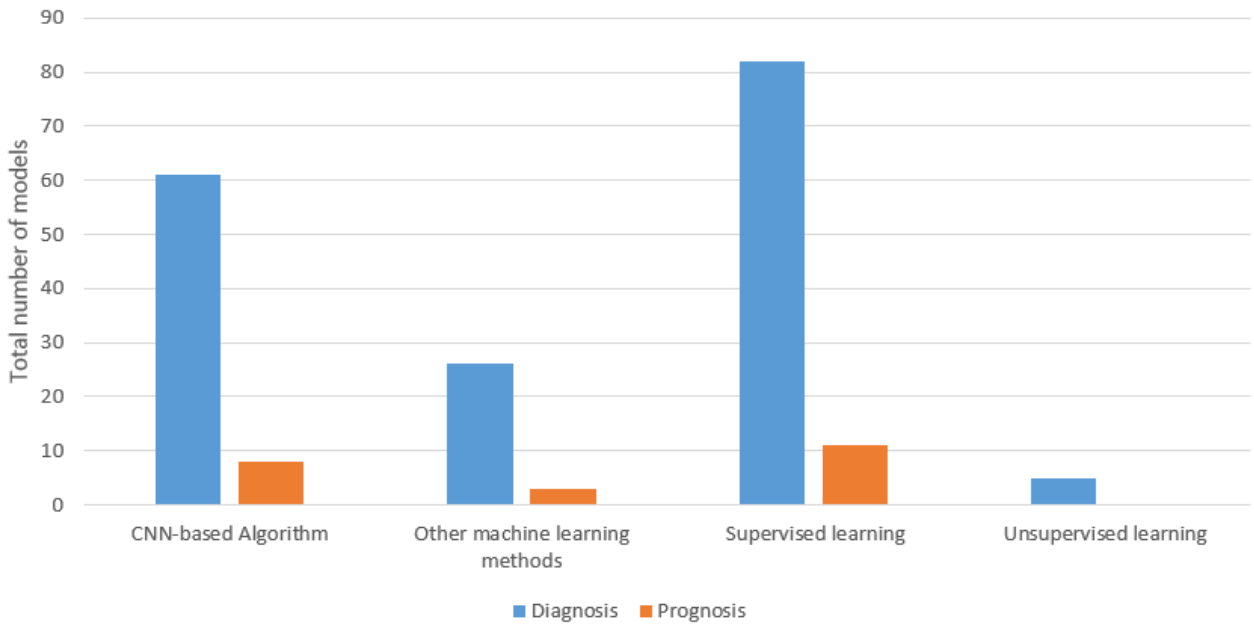
and externally validated in the same study (in an independent dataset, excluding random training test splits and temporal splits) [28,30,46,55].

Prognostic Models for Patients With a COVID-19 Diagnosis

We identified 4 prognostic models for patients who had received a COVID-19 diagnosis. One of these models used a CNN-based model to estimate mortality risk in patients with suspected or confirmed COVID-19 and externally validate using another dataset [43]. Two models aimed to predict disease progression to a severe or critical state, and one of these two models used five CNN-based algorithms [41]. The fourth prognostic model used an LSTM network and compared it with other traditional methods such as principal component analysis, linear discriminant analysis, SVM, and multilayer perceptron [71]. Furthermore, 1 study [72] aimed to develop a random forest algorithm and a logistic regression model to predict the length of hospital stay (greater than 10 days) and estimated C indices of 0.92 and 0.96, respectively. The other studies did not report the C index. Figure 2 shows the bar graph for all methods used in the included studies.

In our analysis, we found that almost all studies had problems with the lack of sufficient data. To address this problem, some studies used data augmentation to synthesize new data, some others attempted to use a combination of different datasets or different kinds of data in their study, and other studies tried to take advantages of non–neural network–based methods such as k-nearest neighbor, SVM, and feature extraction methods. In general, studies that used deep CNNs produced better results than those using non–neural network–based methods. Moreover, 18 studies used K-fold cross-validation, whereas 19 of them used random training test split as a validation method.

Figure 2. Number of deep learning and other machine learning methods used in the reviewed studies. CNN: convolutional neural network.



Risk of Bias

According to the PROBAST assessment tool [26], all included studies were at a high risk of bias, which suggests that their predictive performance when used in practice is probably lower than that reported. Most of the studies were at high risk in the participant domain due to the lack of information about patients and intervention groups. Moreover, almost all studies obtained a high index in the analysis domain, which shows that most of the deep learning models did not have interpretability and that the results were probably lower than those obtained using real datasets.

As shown in Table 2, 15 of the 44 (34%) studies had a high risk of bias for the participant domain, which indicates that these articles did not contain adequate information about the enrolled study participants and intervention groups. In addition, any imbalances in the datasets could cause problems in the randomization process (eg, imbalances between the number of images of normal cases and COVID-19 or other pneumonia cases), leading the study to a risk of bias. Unclear reporting on the inclusion of participants prohibited a risk of bias assessment in 15 (34%) studies. On the other hand, 19 (43%) studies had a high risk of bias due to the predictor domain; this may be attributed to the high false-negative ratio of COVID-19 diagnostic tests (eg, RT-PCR) due to which CT and x-ray images may be wrongly classified as COVID-19, thus leading to inaccurate learning of the models and missing outcome data to predicting processes. In addition, an unclear index was reported in 13 (30%) articles, implying that these articles did not provide specific information about the missing outcome data.

Table 2. Risk of bias assessment (using Prediction Model Risk of Bias Assessment Tool) based on four domains conducted for all studies included in the review.

Study	Domain				Overall risk of bias
	Participants	Predictors	Outcome	Analysis	
[27]	Unclear	Low	High	Unclear	High
[28]	High	High	High	High	High
[30]	Unclear	Unclear	High	High	High
[32]	Unclear	High	High	High	High
[34]	High	Unclear	Unclear	High	High
[73]	Unclear	Unclear	Unclear	High	High
[35]	High	High	High	High	High
[36]	High	High	Low	Unclear	High
[37]	Low	Low	Unclear	Unclear	Unclear
[38]	Unclear	High	Low	High	High
[39]	High	Unclear	Unclear	High	Unclear
[40]	Unclear	Low	Low	High	High
[41]	Low	Low	Low	High	High
[42]	Some concern	High	Low	Unclear	High
[76]	High	High	High	High	High
[43]	Unclear	Low	High	Unclear	High
[45]	High	High	Low	High	High
[46]	High	High	High	High	High
[75]	Unclear	High	Unclear	High	High
[74]	Unclear	Low	Unclear	Unclear	Unclear
[47]	Unclear	Low	High	Unclear	High
[48]	Unclear	Low	Unclear	High	High
[49]	Low	Unclear	Low	High	High
[77]	Unclear	High	High	High	High
[50]	Low	High	High	High	High
[51]	Unclear	High	High	High	High
[52]	High	High	High	High	High
[53]	High	High	Low	High	High
[54]	High	High	High	High	High
[55]	High	High	Low	High	High
[56]	Low	Low	Unclear	High	High
[57]	Unclear	Low	High	High	High
[58]	High	High	High	High	High
[60]	High	High	High	High	High
[61]	High	Unclear	Low	High	High
[63]	High	Unclear	High	High	High
[64]	Unclear	Unclear	High	High	High
[65]	High	Unclear	Low	High	High
[66]	High	Unclear	Low	High	High
[67]	High	Unclear	High	High	High

Study	Domain				Overall risk of bias
	Participants	Predictors	Outcome	Analysis	
[69]	Unclear	Unclear	Low	High	High
[70]	Unclear	Unclear	High	High	High
[71]	Low	Unclear	Unclear	High	High
[72]	Unclear	Unclear	Low	Low	High

Published research articles often do not provide clear information about the preprocessing steps, such as cropping of images. Furthermore, due to the complexity of the machine learning algorithms used to process images into predictors, it is challenging to fully apply the PROBAST predictors. Most models were at high risk of bias in the outcome domain because most of the studies used inappropriate measurement, or there was no reason that the measurement or ascertainment of the outcome differed among intervention groups. Finally, none of the models were identified to be at low risk of bias in the analysis domain. Although many datasets have been made available to researchers in recent months to diagnose COVID-19, there remains a lack of training data, which increases the risk of overfitting. Five models were developed and externally validated in the same study (in an independent dataset, excluding random training test splits and temporal splits).

Metrics

For a more comprehensive review, we classified machine learning–based COVID-19 diagnostic techniques into three major categories based on the imaging modality used in the study. In the following sections, we discuss each category in detail.

CT Scan Category

all machine learning methods that were classified in the CT category used CT scan images in their analyses. Since CT scan data have a 3D nature, two approaches were generally followed. The first is a slice-based approach in which each slice of a CT scan image is analyzed independently; then, at the stage of decision-making, voting is used to decide whether the CT scan image belongs to COVID-19–positive class or COVID-19–negative class. In the second approach, all slices of a CT scan were used as a 3D-like set and used in a 3D CNN [45,57]. The investigations showed that methods utilizing a slice-based approach have a better performance in terms of COVID-19 diagnosis.

For example, Pu et al [45] proposed three 3D CNN models to classify pneumonia and COVID-19 cases by using CT scans. They analyzed 498 CT scans of patients with COVID-19 and 497 CT scans of patients with pneumonia in their experiments. Thus, 256 slices of each CT scan were used as input to the models. Although the results showed that the model with a higher number of layers had the best performance with an area under the curve (AUC) of 0.7, their model could not distinguish between pneumonia and COVID-19 well enough.

Among the methods utilizing a slice-based approach, the proposed method by Ardakani et al [37] reported the best performance with an accuracy of 0.99 and a sensitivity of 1.0. They trained 10 different well-known CNNs by using 1020 slices of 108 CT scans to distinguish COVID-19 from other pneumonias and normal cases. ResNet-101 demonstrated the best sensitivity and was reported as an efficient model for COVID-19 diagnosis by using CT images. Although ResNet-101 had the best sensitivity, it had the weakest results in terms of specificity as compared to Xception and ResNet-50 models, which implies that ResNet-101 might be involved in overfitting.

Some other studies [28,41,56] also reported an accuracy higher than 0.96. The common factor in these approaches was the high level of augmentation used. For instance, Zhang et al [41] used 4695 CT slices that was increased to more than 600,000 slices by using augmentation techniques. Owing to the significance of the number of available images in the training of deep CNN models, some studies attempted to use non–CNN-based methods such as feature extraction, thresholding, and transformation-based methods.

As an example, Fang et al [74] used a radiometric feature extraction technique for all slices of available CT scans (including CT scans of 46 COVID-19–positive and 26 other pneumonia cases); the extracted features were used to train an SVM classifier for further classification. In the test phase, their method achieved an AUC of 0.76. Because other measurements such as accuracy and sensitivity were not reported [74], high risk of bias is very probable.

Due to the difference in color and texture of healthy and infected regions in the lung images, some researchers tried to exploit texture information in their studies. For example, El Asnaoui et al [40] used different feature descriptors such as local binary pattern, gray level co-occurrence matrix, and discrete wavelet transform to analyze local features in images. Finally, in the decision-making stage, an SVM classifier was used to determine whether an input image belongs to the COVID-19 class or not. The results show that this method could achieve a sensitivity of 0.93 and a specificity of 1.0.

X-ray Category

Although a CT scan generates high-quality images with more details than an x-ray image, some studies have attempted to use x-ray images to investigate the probability of COVID-19 diagnosis. Among the studies we reviewed, 14 studies used CXR images in their analyses. Yi et al [46] proposed a hypothesis that a deep CNN model trained on a similar dataset can be useful in COVID-19 diagnosis. They trained a ResNet model for pulmonary tuberculosis (TB) detection by using CXR images from the NIH Chest X-ray dataset [78], which did not

have any information of TB, yet the trained model achieved a high performance with regard to TB detection. The same approach had been used for COVID-19 diagnosis, and the x-ray images of 88 COVID-19–positive patients were inputted into the trained model. The results showed that the model could correctly classify 78 of the 88 (89%) input x-ray images and that it misclassified 10 input x-ray images. Although the reported results are satisfactory, they did not consider COVID-19–negative inputs and did not measure the false-positive rates of the proposed methods.

A continuously growing dataset has been provided by a group of researchers at the University of Montreal [79], which includes annotated CXR images of patients with COVID-19. Several studies [39,40,47,48,51,55] have used this dataset in their analyses. For instance, Han et al [55] proposed a DenseNet model with a relatively small number of parameters and used a combination of x-ray images from various datasets, including the COVID Chest X-Ray dataset (180 COVID-19–positive images), JSRT (20 normal images), NLM (73 normal and 57 tuberculosis images), and CoronaHack (98 normal and 54 pneumonia images), for the training and testing phases. The trained model achieved an accuracy of 0.88 and a precision of 0.83.

Another study [27] utilized images from a pneumonia dataset, including 22,000 CXR images, to train a U-Net model to compute the probability of pneumonia using x-ray images at the pixel level. By integrating the probability values of pixels as a single image, a class activation map is obtained that can be used to show which region in the input image has the most relevance to pneumonia. After model training, they fed 10 CXR images from 5 patients that were captured on several consecutive days. They reported that their model could detect localized areas of pneumonia with increasing likelihood as the subtle airspace opacities increased over time. However, no technical information and measurements were described.

Some other studies [35,48,55] also used a class activation map to not only classify each image into COVID-19–positive and COVID-19–negative classes but also to localize suspected areas in CXR images.

Hybrid Category

Given that most of the included articles mentioned data shortage as a major problem in developing an efficient COVID-19 diagnosis model, some studies tried to exploit two or more types of data in their analyses. For instance, in the study by Wang et al [43], at the first stage, a CNN model was trained on 4106 CT slices with epidermal growth factor receptor data. In the second stage, 709 COVID-19–positive images from patients from Wuhan city were used to retrain the model. Finally, 458 images from four different cities in China were used as test images, and the model achieved an accuracy of 0.85 and a sensitivity of 0.80.

In the study by Mei et al [50], clinical data such as patient's age, gender, symptoms, and laboratory findings were used in addition to CT scans of 905 patients with suspected COVID-19 from 13 provinces in China. A modified ResNet model was proposed by the authors to accept clinical data alongside the CT scan slice images. The results showed that their proposed model achieved an accuracy equivalent to a senior chest radiologist with an AUC of 0.86. Although their dataset is not publicly available, the trained models are available for others to download.

Discussion

Principal Findings

In this study, we reviewed 44 studies related to the diagnosis and prognosis of COVID-19 that used advanced machine learning techniques based on clinical images to diagnose COVID-19 or COVID-19–related pneumonia, or to assist with the segmentation of lung images by using chest CT and x-ray images with their proposed machine learning methods. The predictive performance measures showed a high to almost perfect ability to detect COVID-19. Overall, 24 different methods, such as deep CNNs, local feature descriptors, and decision trees, were used in the reviewed studies; however, some of them used similar models with a different setup or configuration.

Due to the complexity of the clinical images used and the need to obtain the best results for an early diagnosis of COVID-19, most of the reviewed articles (36/44, 82%) had based their learning algorithm on neural networks and deep learning as proven, powerful learning methods. However, deep CNNs, which are developed in principle to work with images, require sufficient amount of data for fine-tuning the network parameters.

Given that the COVID-19 outbreak was in the early stage at the time of this review and that there was a lack of proper data available, most of these CNN-based studies were endangered by overfitting, which causes a high risk of bias. Nevertheless, some of the studies used previously available data of chest CT or x-ray images to compensate with data shortage and to enrich the training data. For instance, Ucar and Korkmaz [38] used 66 COVID-19–positive lung x-ray images, which were not sufficient to train a CNN. To overcome this problem, they added these images to the images of a publicly available pneumonia dataset called Chest X-Ray Images (Pneumonia) [80], which was used to obtain access to a larger number of images for network training. Although the pneumonia dataset does not provide any information about COVID-19, it can enhance the model performance to better distinguish between healthy and unhealthy lungs. Another approach used for compensating the lack of data was to utilize data augmentation techniques such as image mirroring and blending. Although most of the reviewed studies used simple augmentation methods, some used more complicated techniques. For example, in the study by Ucar and Korkmaz [38], a generative adversarial network was trained to synthesize new images from the limited 307 images available that were not considered enough for network training.

This systematic review is in its early stage, and we will continue to update our findings and evaluation to provide new information to health care professionals and decision makers as more international studies are conducted over time.

Study Limitations

With the rapid publication of COVID-19 prediction models in the medical image processing domain in the recent past, this systematic review cannot be considered as an up-to-date list of all the current prediction models.

Conclusions

Different models have been proposed for the diagnosis and prognosis of COVID-19, demonstrating varying levels of discriminative performance. The results show that deep CNNs dedicated a larger number of models than non–neural network–based methods; moreover, deep networks achieved better results than other machine learning models. However, the rapid spread of COVID-19 and the lack of data for machine learning approaches and training may have increased the likelihood of overfitting and vague reporting. Furthermore, the lack of adequate information about patients and study participants likely led to the high risk of bias, which made the results seem optimistic. Therefore, the performance of these models is misleading, and we do not recommend their practical use. Future studies aimed at using deep neural networks for diagnosing COVID-19 should address aspects of appropriate model performance by using a larger training dataset with no imbalance and complete information about patients and intervention groups.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Table S1. Characteristics of included articles.
[PDF File (Adobe PDF File), 227 KB - medinform_v9i4e25181_app1.pdf]

References

1. Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [accessed 2021-03-30]
2. Rolling updates on coronavirus disease (COVID-19). World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen [accessed 2021-03-30]
3. Sohrabi C, Alsafi Z, O'Neill N, Khan M, Kerwan A, Al-Jabir A, et al. World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). Int J Surg 2020 Apr;76:71-76 [FREE Full text] [doi: 10.1016/j.ijsu.2020.02.034] [Medline: 32112977]
4. Arabi YM, Murthy S, Webb S. COVID-19: a novel coronavirus and a novel challenge for critical care. Intensive Care Med 2020 May;46(5):833-836 [FREE Full text] [doi: 10.1007/s00134-020-05955-1] [Medline: 32125458]
5. Xie J, Tong Z, Guan X, Du B, Qiu H, Slutsky AS. Critical care crisis and some recommendations during the COVID-19 epidemic in China. Intensive Care Med 2020 May;46(5):837-840 [FREE Full text] [doi: 10.1007/s00134-020-05979-7] [Medline: 32123994]
6. World Health Organization. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance [accessed 2021-03-30]
7. Fang Y, Zhang H, Xie J, Lin M, Ying L, Pang P, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology 2020 Aug;296(2):E115-E117 [FREE Full text] [doi: 10.1148/radiol.2020200432] [Medline: 32073353]
8. Huang L, Han R, Yu P, Wang S, Xia L. A correlation study of CT and clinical features of different clinical types of COVID-19. Chinese J Radiol 2020 Apr 10;54(4):304. [doi: 10.3760/cma.j.cn112149-20200205-00087]
9. Park JH, LOCAT Group. Diagnostic imaging utilization in cases of acute appendicitis: multi-center experience. J Korean Med Sci 2014 Sep;29(9):1308-1316 [FREE Full text] [doi: 10.3346/jkms.2014.29.9.1308] [Medline: 25246752]
10. Chung M, Bernheim A, Mei X, Zhang N, Huang M, Zeng X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology 2020 Apr;295(1):202-207. [doi: 10.1148/radiol.2020200230] [Medline: 32017661]
11. Yang X, Yu Y, Xu J, Shu H, Xia J, Liu H, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med 2020 May;8(5):475-481 [FREE Full text] [doi: 10.1016/S2213-2600(20)30079-5] [Medline: 32105632]
12. Kanne JP. Chest CT findings in 2019 novel coronavirus (2019-nCoV) infections from Wuhan, China: Key points for the radiologist. Radiology 2020 Apr;295(1):16-17. [doi: 10.1148/radiol.2020200241] [Medline: 32017662]
13. Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection. Radiology 2020 Jun;295(3):200463 [FREE Full text] [doi: 10.1148/radiol.2020200463] [Medline: 32077789]
14. Kay FU, Abbara S. The many faces of COVID-19: Spectrum of imaging manifestations. Radiology: Cardiothoracic Imaging 2020 Feb 01;2(1):e200037. [doi: 10.1148/ryct.2020200037]
15. Ng M, Lee E, Yang J, Yang F, Li X, Wang H, et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. Radiol Cardiothorac Imaging 2020 Mar;2(1):e200034. [doi: 10.1148/ryct.2020200034] [Medline: 33778547]

16. Kong B, Wang X, Bai J, Lu Y, Gao F, Cao K, et al. Learning tree-structured representation for 3D coronary artery segmentation. Comput Med Imaging Graph 2020 Mar;80:101688. [doi: 10.1016/j.compmedimag.2019.101688] [Medline: 31926366]
17. Xia C, Li X, Wang X, Kong B, Chen Y, Yin Y. A multi-modality network for cardiomyopathy death risk prediction with CMR images and clinical information. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). Springer.? 2019:585-577. [doi: 10.1007/978-3-030-32245-8_64]
18. Ramani V, Shendure J. Smash and DASH with Cas9. Genome Biol 2016 Mar 05;17:42 [FREE Full text] [doi: 10.1186/s13059-016-0905-4] [Medline: 26944856]
19. Rajaraman S, Candemir S, Kim I, Thoma G, Antani S. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl Sci (Basel) 2018 Oct;8(10):1715 [FREE Full text] [doi: 10.3390/app8101715] [Medline: 32457819]
20. Depeursinge A, Chin AS, Leung AN, Terrone D, Bristow M, Rosen G, et al. Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution computed tomography. Invest Radiol 2015 Apr;50(4):261-267 [FREE Full text] [doi: 10.1097/RLI.0000000000000127] [Medline: 25551822]
21. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans Med Imaging 2016 May;35(5):1207-1216. [doi: 10.1109/TMI.2016.2535865] [Medline: 26955021]
22. Sharing research data and findings relevant to the novel coronavirus (COVID-19) outbreak. Wellcome Trust. 2020 Jan 31. URL: https://wellcome.org/coronavirus-covid-19/open-data [accessed 2021-03-31]
23. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, PRISMA-IPD Development Group. Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA 2015 Apr 28;313(16):1657-1665. [doi: 10.1001/jama.2015.3656] [Medline: 25919529]
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev 2016 Dec 05;5(1):210 [FREE Full text] [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
25. O'Caoimh R, Cornally N, Weathers E, O'Sullivan R, Fitzgerald C, Orfila F, et al. Risk prediction in the community: A systematic review of case-finding instruments that predict adverse healthcare outcomes in community-dwelling older adults. Maturitas 2015 Sep;82(1):3-21 [FREE Full text] [doi: 10.1016/j.maturitas.2015.03.009] [Medline: 25866212]
26. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: A tool to assess risk of bias and applicability of prediction model studies: Explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):W1. [doi: 10.7326/m18-1377]
27. Hurt B, Kligerman S, Hsiao A. Deep learning localization of pneumonia: 2019 coronavirus (COVID-19) outbreak. J Thorac Imaging 2020 May;35(3):W87-W89 [FREE Full text] [doi: 10.1097/RTI.0000000000000512] [Medline: 32205822]
28. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 2020 Sep;296(3):E156-E165 [FREE Full text] [doi: 10.1148/radiol.2020201491] [Medline: 32339081]
29. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. 2019 Presented at: 36th International Conference on Machine Learning, ICML 2019; June 2019; Long Beach, CA p. 10691-10700.
30. Gozes O, Frid-Adar M, Greenspan H, Browning P, Zhang H, Ji W. Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis. arXiv. Preprint posted online on March 10, 2020 [FREE Full text]
31. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV p. 770. [doi: 10.1109/cvpr.2016.90]
32. Kang H, Xia L, Yan F, Wan Z, Shi F, Yuan H, et al. Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. IEEE Trans Med Imaging 2020 Aug;39(8):2606-2614. [doi: 10.1109/TMI.2020.2992546] [Medline: 32386147]
33. Zhang C, Han Z, Cui Y, Fu H, Zhou JT, Hu Q. CPM-Nets: Cross partial multi-view networks. In: Wallach H, Larochelle H, Beygelzimer A, textquotesingle Alch F, Buc E, Fox E, et al, editors. Advances in Neural Information Processing Systems. 57 Morehouse Ln, Red Hook, NY 12571: Curran Associates, Inc; 2019.
34. Jin C, Chen W, Cao Y, Xu Z, Tan Z, Zhang X, et al. Development and evaluation of an artificial intelligence system for COVID-19 diagnosis. Nat Commun 2020 Oct 09;11(1):5088 [FREE Full text] [doi: 10.1038/s41467-020-18685-1] [Medline: 33037212]
35. Murphy K, Smits H, Knoops AJG, Korst MBJM, Samson T, Scholten ET, et al. COVID-19 on chest radiographs: A multireader evaluation of an artificial intelligence system. Radiology 2020 Sep;296(3):E166-E172. [doi: 10.1148/radiol.2020201874] [Medline: 32384019]
36. Loey M, Smarandache F, Khalifa NEM. Within the lack of chest COVID-19 X-ray dataset: A novel detection model based on GAN and deep transfer learning. Symmetry 2020 Apr 20;12(4):651. [doi: 10.3390/sym12040651]
37. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med 2020 Jun;121:103795 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103795] [Medline: 32568676]

38. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images. Med Hypotheses 2020 Jul;140:109761 [FREE Full text] [doi: 10.1016/j.mehy.2020.109761] [Medline: 32344309]
39. Vaid S, Kalantar R, Bhandari M. Deep learning COVID-19 detection bias: accuracy through artificial intelligence. Int Orthop 2020 Aug;44(8):1539-1542 [FREE Full text] [doi: 10.1007/s00264-020-04609-7] [Medline: 32462314]
40. El Asnaoui K, Chawki Y. Using X-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn 2020 May 22:1-12 [FREE Full text] [doi: 10.1080/07391102.2020.1767212] [Medline: 32397844]
41. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Jun 11;181(6):1423-1433.e11 [FREE Full text] [doi: 10.1016/j.cell.2020.04.045] [Medline: 32416069]
42. Wu X, Hui H, Niu M, Li L, Wang L, He B, et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. Eur J Radiol 2020 Jul;128:109041 [FREE Full text] [doi: 10.1016/j.ejrad.2020.109041] [Medline: 32408222]
43. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020 Aug;56(2):2000775 [FREE Full text] [doi: 10.1183/13993003.00775-2020] [Medline: 32444412]
44. Huang G, Liu Z, Van DML, Weinberger K. Densely connected convolutional networks. 2017 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI p. 2261-2269. [doi: 10.1109/cvpr.2017.243]
45. Pu J, Leader J, Bandos A, Shi J, Du P, Yu J, et al. Any unique image biomarkers associated with COVID-19? Eur Radiol 2020 Jul 20;30(11):6221-6227. [doi: 10.1007/s00330-020-06956-w] [Medline: 32462445]
46. Yi PH, Kim TK, Lin CT. Generalizability of deep learning tuberculosis classifier to COVID-19 chest radiographs: New tricks for an old algorithm? J Thorac Imaging 2020 Jul;35(4):W102-W104. [doi: 10.1097/RTI.0000000000000532] [Medline: 32427650]
47. Toğaçar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest X-ray images using fuzzy color and stacking approaches. Comput Biol Med 2020 Jun;121:103805 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103805] [Medline: 32568679]
48. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Rajendra Acharya U. Automated detection of COVID-19 cases using deep neural networks with X-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]
49. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. Radiology 2020 Aug;296(2):E65-E71 [FREE Full text] [doi: 10.1148/radiol.2020200905] [Medline: 32191588]
50. Mei X, Lee H, Diao K, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020 Aug;26(8):1224-1228 [FREE Full text] [doi: 10.1038/s41591-020-0931-3] [Medline: 32427924]
51. Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting possibly representative COVID-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. J Med Biol Eng 2020 May 14:1-8 [FREE Full text] [doi: 10.1007/s40846-020-00529-4] [Medline: 32412551]
52. Yang S, Jiang L, Cao Z, Wang L, Cao J, Feng R, et al. Deep learning for detecting corona virus disease 2019 (COVID-19) on high-resolution computed tomography: a pilot study. Ann Transl Med 2020 Apr;8(7):450 [FREE Full text] [doi: 10.21037/atm.2020.03.132] [Medline: 32395494]
53. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. IEEE Access 2020;8:91916-91923. [doi: 10.1109/access.2020.2994762]
54. Wang X, Deng X, Fu Q, Zhou Q, Feng J, Ma H, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans Med Imaging 2020 Aug;39(8):2615-2625. [doi: 10.1109/TMI.2020.2995965] [Medline: 33156775]
55. Oh Y, Park S, Ye JC. Deep learning COVID-19 features on CXR using limited training data sets. IEEE Trans Med Imaging 2020 Aug;39(8):2688-2700. [doi: 10.1109/TMI.2020.2993291] [Medline: 32396075]
56. Han Z, Wei B, Hong Y, Li T, Cong J, Zhu X, et al. Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning. IEEE Trans Med Imaging 2020 Aug;39(8):2584-2594. [doi: 10.1109/TMI.2020.2996256] [Medline: 32730211]
57. Wang J, Bao Y, Wen Y, Lu H, Luo H, Xiang Y, et al. Prior-attention residual learning for more discriminative COVID-19 screening in CT images. IEEE Trans Med Imaging 2020 Aug;39(8):2572-2583. [doi: 10.1109/TMI.2020.2994908] [Medline: 32730210]
58. Roy S, Menapace W, Oei S, Luijten B, Fini E, Saltori C, et al. Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound. IEEE Trans Med Imaging 2020 Aug;39(8):2676-2687. [doi: 10.1109/TMI.2020.2994459] [Medline: 32406829]
59. van Sloun RJG, Demi L. Localizing B-lines in lung ultrasonography by weakly supervised deep learning, in-vivo results. IEEE J Biomed Health Inform 2020 Apr;24(4):957-964. [doi: 10.1109/JBHI.2019.2936151] [Medline: 31425126]

60. Fan D, Zhou T, Ji G, Zhou Y, Chen G, Fu H, et al. Inf-Net: Automatic COVID-19 lung infection segmentation from CT images. IEEE Trans Med Imaging 2020 Aug;39(8):2626-2637. [doi: 10.1109/TMI.2020.2996645] [Medline: 32730213]
61. Chen J, Wu L, Zhang J, Zhang L, Gong D, Zhao Y, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. Sci Rep 2020 Nov 05;10(1):19196 [FREE Full text] [doi: 10.1038/s41598-020-76282-0] [Medline: 33154542]
62. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: A nested U-Net architecture for medical image segmentation. Deep Learn Med Image Anal Multimodal Learn Clin Decis Support (2018) 2018 Sep;11045:3-11 [FREE Full text] [doi: 10.1007/978-3-030-00889-5_1] [Medline: 32613207]
63. Wang B, Jin S, Yan Q, Xu H, Luo C, Wei L, et al. AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system. Appl Soft Comput 2021 Jan;98:106897 [FREE Full text] [doi: 10.1016/j.asoc.2020.106897] [Medline: 33199977]
64. Shan F, Gao Y, Wang J, Shi W, Shi N, Han M. Lung infection quantification of COVID-19 in CT images with deep learning. arXiv. Preprint posted online on March 10, 2020.
65. Shi F, Xia L, Shan F, Wu D, Wei Y, Yuan H. Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification. arXiv. Preprint posted online on March 22, 2020.
66. Wang S, Kang B, Ma J, Zeng X, Xiao M, Guo J. A deep learning algorithm using CT images to screen for corona virus disease (COVID-19). medRxiv. Preprint posted online on April 24, 2020. [doi: 10.1101/2020.02.14.20023028]
67. Xu X, Jiang X, Ma C, Du P, Li X, Lv S, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. Engineering (Beijing) 2020 Oct;6(10):1122-1129 [FREE Full text] [doi: 10.1016/j.eng.2020.04.010] [Medline: 32837749]
68. Li X, Zhou Y, Du P, Lang G, Xu M, Wu W. A deep learning system that generates quantitative CT reports for diagnosing pulmonary tuberculosis. Applied Intelligence 2020 Nov 26:1-12. [doi: 10.1007/s10489-020-02051-1]
69. Song Y, Zheng S, Li L, Zhang X, Zhang X, Huang Z, et al. Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images. IEEE/ACM Trans Comput Biol Bioinform 2021 Mar 11. [doi: 10.1109/TCBB.2021.3065361] [Medline: 33705321]
70. Wang X, Deng X, Fu Q, Zhou Q, Feng J, Ma H, et al. A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. IEEE Trans Med Imaging 2020 Aug;39(8):2615-2625. [doi: 10.1109/TMI.2020.2995965] [Medline: 33156775]
71. Bai X, Fang C, Zhou Y, Bai S, Liu Z, Xia L, et al. Predicting COVID-19 malignant progression with AI techniques. SSRN. Preprint posted online on March 31, 2020. [doi: 10.2139/ssrn.3557984]
72. Yue H, Yu Q, Liu C, Huang Y, Jiang Z, Shao C, et al. Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study. Ann Transl Med 2020 Jul;8(14):859 [FREE Full text] [doi: 10.21037/atm-20-3026] [Medline: 32793703]
73. Barstugan M, Ozkaya U, Ozturk S. Coronavirus (COVID-19) classification using CT images by machine learning methods. arXiv. Preprint posted online on March 20, 2020.
74. Fang M, He B, Li L, Dong D, Yang X, Li C, et al. CT radiomics can help screen the coronavirus disease 2019 (COVID-19): a preliminary study. Sci China Inf Sci 2020 Apr 15;63(7):1-8. [doi: 10.1007/s11432-020-2849-3]
75. Tuncer T, Dogan S, Ozyurt F. An automated Residual Exemplar Local Binary Pattern and iterative ReliefF based COVID-19 detection method using chest X-ray image. Chemometr Intell Lab Syst 2020 Aug 15;203:104054 [FREE Full text] [doi: 10.1016/j.chemolab.2020.104054] [Medline: 32427226]
76. Pereira RM, Bertolini D, Teixeira LO, Silla CN, Costa YMG. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed 2020 Oct;194:105532 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105532] [Medline: 32446037]
77. Singh D, Kumar V, Vaishali, Kaur M. Classification of COVID-19 patients from chest CT images using multi-objective differential evolution-based convolutional neural networks. Eur J Clin Microbiol Infect Dis 2020 Jul;39(7):1379-1389 [FREE Full text] [doi: 10.1007/s10096-020-03901-z] [Medline: 32337662]
78. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Xplore. 2017 May 05 Presented at: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI URL: https://ieeexplore.ieee.org/document/8099852 [doi: 10.1109/CVPR.2017.369]
79. 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]
80. Rahman T. COVID-19 Radiography Database. Kaggle. URL: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database [accessed 2021-04-05]

Abbreviations

AUC: area under the curve
CNN: convolutional neural network

CT: computed tomography
CXR: chest x-ray
GAN: generative adversarial network
PROBAST: Prediction Model Risk of Bias Assessment Tool
RT-PCR: reverse transcription–polymerase chain reaction
SVM: support vector machine

Edited by C Lovis; submitted 21.10.20; peer-reviewed by F Khorami, A Drory; comments to author 07.11.20; revised version received 31.12.20; accepted 16.01.21; published 23.04.21.

Please cite as:
Montazeri M, ZahediNasab R, Farahani A, Mohseni H, Ghasemian F
Machine Learning Models for Image-Based Diagnosis and Prognosis of COVID-19: Systematic Review
JMIR Med Inform 2021;9(4):e25181
URL: https://medinform.jmir.org/2021/4/e25181
doi:10.2196/25181
PMID:33735095

Original Paper

Application of Artificial Intelligence for Screening COVID-19 Patients Using Digital Images: Meta-analysis

Tahmina Nasrin Poly[1,2,3], MSc; Md Mohaimenul Islam[1,2,3], MSc, PhD; Yu-Chuan Jack Li[1,2,3,4,5], MD, PhD; Belal Alsinglawi[6], MSc; Min-Huei Hsu[7], PhD; Wen Shan Jian[8], PhD; Hsuan-Chia Yang[1,2,3], PhD

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[2]International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan

[3]Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[4]Department of Dermatology, Wan Fang Hospital, Taipei Medical University, Taipei, Taiwan

[5]TMU Research Center of Cancer Translational Medicine, Taipei Medical University, Taipei, Taiwan

[6]School of Computer, Data, and Mathematical Science, Western Sydney University, Sydney, Australia

[7]Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan

[8]School of Health Care Administration, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Hsuan-Chia Yang, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
15 Floor, No. 172-1, Section: 2, Keelung Road, Daan District
Taipei, 106
Taiwan
Phone: 886 (02)66382736 ext 1507
Email: itpharmacist@gmail.com

Abstract

Background: The COVID-19 outbreak has spread rapidly and hospitals are overwhelmed with COVID-19 patients. While analysis of nasal and throat swabs from patients is the main way to detect COVID-19, analyzing chest images could offer an alternative method to hospitals, where health care personnel and testing kits are scarce. Deep learning (DL), in particular, has shown impressive levels of performance when analyzing medical images, including those related to COVID-19 pneumonia.

Objective: The goal of this study was to perform a systematic review with a meta-analysis of relevant studies to quantify the performance of DL algorithms in the automatic stratification of COVID-19 patients using chest images.

Methods: A search strategy for use in PubMed, Scopus, Google Scholar, and Web of Science was developed, where we searched for articles published between January 1 and April 25, 2020. We used the key terms “COVID-19,” or “coronavirus,” or “SARS-CoV-2,” or “novel corona,” or “2019-ncov,” and “deep learning,” or “artificial intelligence,” or “automatic detection.” Two authors independently extracted data on study characteristics, methods, risk of bias, and outcomes. Any disagreement between them was resolved by consensus.

Results: A total of 16 studies were included in the meta-analysis, which included 5896 chest images from COVID-19 patients. The pooled sensitivity and specificity of the DL models in detecting COVID-19 were 0.95 (95% CI 0.94-0.95) and 0.96 (95% CI 0.96-0.97), respectively, with an area under the receiver operating characteristic curve of 0.98. The positive likelihood, negative likelihood, and diagnostic odds ratio were 19.02 (95% CI 12.83-28.19), 0.06 (95% CI 0.04-0.10), and 368.07 (95% CI 162.30-834.75), respectively. The pooled sensitivity and specificity for distinguishing other types of pneumonia from COVID-19 were 0.93 (95% CI 0.92-0.94) and 0.95 (95% CI 0.94-0.95), respectively. The performance of radiologists in detecting COVID-19 was lower than that of the DL models; however, the performance of junior radiologists was improved when they used DL-based prediction tools.

Conclusions: Our study findings show that DL models have immense potential in accurately stratifying COVID-19 patients and in correctly differentiating them from patients with other types of pneumonia and normal patients. Implementation of DL-based tools can assist radiologists in correctly and quickly detecting COVID-19 and, consequently, in combating the COVID-19 pandemic.

(JMIR Med Inform 2021;9(4):e21394) doi:10.2196/21394

KEYWORDS

COVID-19; SARS-CoV-2; pneumonia; artificial intelligence; deep learning

Introduction

COVID-19 is a serious global infectious disease and is spreading at an unprecedented level worldwide [1,2]. The World Health Organization declared this infectious disease a public health emergency of international concern and then declared it a pandemic. SARS-CoV-2 is even more contagious than SARS-CoV or Middle East respiratory syndrome coronavirus and is sometimes undetected due to people having asymptomatic or mild symptoms [3,4]. Earlier detection paired with aggressive public health steps, such as social distancing and isolation of suspected or sick patients, can help tackle the crisis [5]. Presently, reverse transcription–polymerase chain reaction (RT-PCR), gene sequencing, and analysis of blood specimens are considered the gold standard methods for detecting COVID-19; however, the performance of these methods (~73% sensitivity for nasal swabs and ~61% for throat swabs) is not satisfactory [6,7]. Since hospitals are overwhelmed by COVID-19 patients, those with severe acute respiratory illness are given priority over others with mild symptoms. Therefore, a large number of undiagnosed patients may lead to a serious risk of cross-infection.

Chest radiography imaging (eg, x-ray and computed tomography [CT] scan) is often used as an effective tool for the quick diagnosis of pneumonia [8,9]. The CT scan images of COVID-19 patients show multilobar involvement and peripheral airspace, mostly ground-glass opacities [10,11]. Moreover, asymmetric patchy or diffuse airspace opacities have also been reported in patients with SARS-CoV-2 infection [12]. These changes in CT scan images can be easily interpreted by a trained or experienced radiologist. Automatic classification of COVID-19 patients, however, has huge benefits, such as increasing efficiency, wide coverage, reducing barriers to access, and improving patient outcomes. Several studies demonstrated the application of deep learning (DL) techniques to identify and detect novel COVID-19 using radiography images [13,14].

Herein, we report the results of a comprehensive systematic review of DL algorithm studies that investigated the performance of DL algorithms for COVID-19 classification from chest radiography imaging. Our main objective was to quantify the performance of DL methods for COVID-19 classification, which might encourage health care policy makers to implement DL-based automated tools in the real-world clinical setting. DL-based automated tools can help reduce radiologists' workload, as DL can help maintain diagnostic radiology support in real time and with increased sensitivity.

Methods

Experimental Approach

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which are based on the Cochrane Handbook for Systematic Reviews of Interventions, were used to conduct this study [15].

Literature Search

We searched electronic databases, such as PubMed, Scopus, Google Scholar, and Web of Science, for articles published between January 1 and April 25, 2020. We developed a search strategy using combinations of the following Medical Subject Headings: "COVID-19," or "coronavirus," or "SARS-CoV-2," or "novel corona," or "2019-ncov," and "deep learning," or "artificial intelligence," or "automatic detection." Reference lists of the retrieved articles and relevant reviews were also checked for additional eligible articles.

Eligibility Criteria

During the first screening, two authors (MMI and TNP) assessed the title and abstract of each article and excluded irrelevant articles. To include eligible articles, those two authors examined the full text of the articles and evaluated whether they fulfilled the inclusion criteria of this study. Disagreement during this selection process was resolved by consensus or, if necessary, the main investigator (YCL) was consulted. We included articles if they met the following criteria: (1) were published in English, (2) were published in a peer-reviewed journal, (3) assessed performance of a DL model to detect COVID-19, and (4) provided a clear description of the methodology and the total number of images. We excluded studies if they were published in preprint repositories or if they were published in the form of a review or a letter to the editor.

Data Extraction and Synthesis

Two authors (MMI and TNP) independently screened all titles and abstracts of retrieved articles. The most relevant studies were selected based on the predefined selection criteria. Any disagreement during the screening process was resolved by discussion with the other authors; unsettled issues were settled by discussion with the study supervisor (YCL). The two authors who conducted the first screening cross-checked studies for duplication by comparing author names, publication dates, and journal names. They excluded all duplicate studies. Afterward, they collected data from the selected studies, such as author name, publication year, location, model description, total number of images, total number of COVID-19 cases and images, imaging modality, total number of patients, sensitivity, specificity, accuracy, area under the receiver operating characteristic curve (AUROC), and database.

Risk of Bias Assessment

The Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool was used to assess the quality of the selected studies [16]. The QUADAS-2 scale comprises four domains: patient selection, index test, reference standard, and flow and timing. The first three domains are used to evaluate the risk of bias in terms of concerns regarding applicability. The overall

risk of bias was categorized into three groups: low, high, and unclear risk of bias.

Statistical Analysis

Meta-DiSc, version 1.4, was used to calculate the evaluation metrics of the DL model. The software was also used to (1) perform statistical pooling from each study and (2) assess the homogeneity with a variety of statistics, including chi-square and I^2. The sensitivity and specificity with 95% CIs in distinguishing between COVID-19 patients, patients with other types of pneumonia, and normal patients were calculated. The pooled receiver operating characteristic (ROC) curve was plotted and the area under the curve (AUC) was calculated with 95% CIs based on the DerSimonian-Laird random effects model method. The diagnostic odds ratio (DOR) was calculated by the Moses constant of the linear model. Diagnostic tests where the DOR is constant, regardless of the diagnostic threshold, have symmetrical curves around the sensitivity-specificity line. In these situations, it is possible to combine DORs using the DerSimonian-Laird method to estimate the overall DOR and, hence, to determine the best-fitting ROC curve [17]. The mathematical equation is given below:

When the DOR changes with the diagnostic threshold, the ROC curve is asymmetrical. To fit the DOR variation based on a different threshold, the Moses-Shapiro-Littenberg method was used. It consists of observing the relationship by fitting the straight line:

$$D = a + bS \quad \textbf{(2)}$$

where D is the log of DOR and S is a measure of threshold given by the following:

Estimates of parameters a and b and their standard errors and covariance were obtained by the ordinary or weighted least squares method using the NAG Library for C (The Numerical Algorithms Group).

The ROC curve is the AUC that summarized the diagnostic performance as a single number: an AUC close to 1 is considered a perfect curve and an AUC close to 0.5 is considered poor [18]. The AUC is computed by numeric integration of the curve equation by the trapezoidal method [19]. The Q^* index is defined by the point where sensitivity and specificity are equal, which is the point closest to the ideal top-left corner of the ROC curve space. It was calculated by the following:

Moreover, the standard error of the AUROC was calculated by following equation:

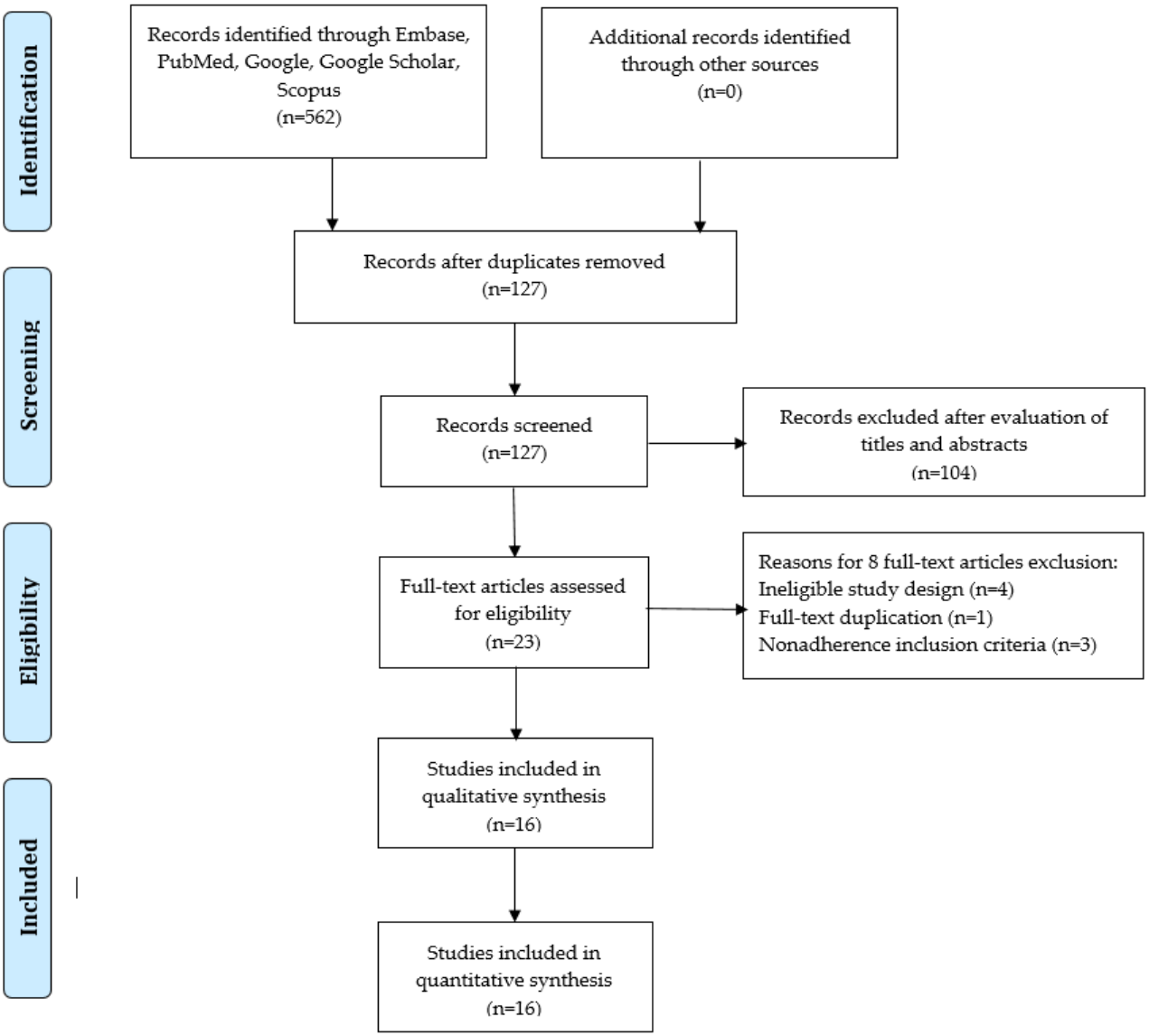
The standard error of Q^* was calculated by following equation:

Results

Selection Criteria

Figure 1 shows the process of identifying relevant DL studies. A total of 562 studies were retrieved by searching electronic databases and by reviewing their reference lists. We excluded 435 duplicate studies and an additional 104 studies that did not fulfill the selection criteria. We reviewed 23 full-text studies and further excluded 7 studies because of the reasons shown in Figure 1. Finally, we included 16 studies in the meta-analysis [13,14,20-33].

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram for study selection.



Characteristics of Included Studies

Among the 16 DL-based COVID-19 detection studies, we identified 5896 digital images for COVID-19 patients and 645,825 images for non-COVID-19 patients, including those with other types of viral pneumonia and normal patients. Included studies used DL algorithms, such as convolutional neural networks, MobileNetV2, and COVNet, for stratifying COVID-19 patients with higher accuracy. The range of accuracy for detecting COVID-19 correctly was 76.00 to 99.51. A total of 8 studies used CT images and 8 studies used x-ray images. The characteristics of the included studies in the meta-analysis are shown in Table 1 [13,14,20-33].

Table 1. Characteristics of the studies included in the meta-analysis.

Author	Modality	Method	Images, n	COVID-19 images, n	Sensitivity	Specificity	Accuracy
Apostolopoulos and Mpesiana [14]	X-ray	MobileNetV2	1428	224	98.66	96.46	99.18
Butt et al [13]	Computed tomography (CT)	Convolutional neural network (CNN)	618	219	98.20	92.20	—[a]
Apostolopoulos et al [21]	X-ray	MobileNetV2	3905	463	97.36	99.42	96.78
Li et al [25]	CT	COVNet	4356	127	90.00	95.00	—
Ucar and Korkmaz [29]	X-ray	CNN	4608[b]	1536[b]	—	99.13	98.30
Ozturk et al [26]	X-ray	CNN and DarkNet	1186	108	95.13	95.30	98.08
Bai et al [24]	CT	EfficientNet	1186	521	95.00	96.00	96.00
Zhang at al [33]	CT	DeepLabv3	617,775	—	94.93	91.13	92.49
El Asnaoui and Chawki [20]	X-ray	Inception ResNet V2	6087	231	92.11	96.06	—
Ardakani et al [22]	CT	ResNet-101	1020	510	100	99.02	99.51
Pathak et al [27]	CT	CNN	852	413	91.45	94.77	93.01
Wu et al [32]	CT	ResNet50	495	368	81.10	61.50	76.00
Toğaçar et al [28]	X-ray	SqueezeNet	458	295	100	100	100
Waheed at al [30]	X-ray	ACGAN[c]	1124	403	90.00	97.00	95.00
Khan et al [23]	X-ray	Xception	1251	284	99.30	98.60	99.00
Wang et al [31]	CT	DenseNet	5372	102	80.39	76.66	78.32
Wang et al [31]	CT	DenseNet	5372	92	79.35	81.16	80.12

[a]Not reported.

[b]Augmented images.

[c]ACGAN: auxiliary classifier generative adversarial network.

Model Performance

Based on the 16 studies, the performance of the DL algorithms for detecting COVID-19 was determined and is summarized in Table 2 [22,24,33]. The pooled sensitivity and specificity of the DL methods for detecting COVID-19 was 0.95 (95% CI 0.94-0.95) and 0.96 (95% CI 0.96-0.97), respectively, with a summary ROC (SROC) of 0.98 (Figure 2). The pooled sensitivity and specificity are shown in Figure 3.

DL methods were able to correctly distinguish other types of pneumonia from COVID-19 with an SROC of 0.98 (sensitivity: 0.93, 95% CI 0.92-0.94; specificity: 0.95, 95% CI 0.94-0.95). The positive likelihood, negative likelihood, and DOR were 22.45 (95% CI 12.86-39.19), 0.06 (95% CI 0.03-0.13), and 461.81 (95% CI 134.96-1580.24), respectively. Moreover, the DL model showed good performance for correctly stratifying normal patients, with an SROC of 0.99 (sensitivity: 0.95, 95% CI 0.94-0.96; specificity: 0.98, 95% CI 0.97-0.98). The positive likelihood, negative likelihood, and DOR were 47.47 (95% CI 20.70-108.86), 0.04 (95% CI 0.02-0.08), and 1524.81 (95% CI 625.29-3718.34), respectively.

Table 2. Performance comparison between deep learning models and radiologists.

Class and method	Data sets, n	Sensitivity (95% CI)	Specificity (95% CI)	Positive likelihood ratio (95% CI)	Negative likelihood ratio (95% CI)	AUROC[a]	Accuracy
COVID-19							
Deep learning model	17	0.95 (0.94-0.95)	0.96 (0.96-0.97)	19.02 (12.83-28.19)	0.06 (0.04-0.10)	0.98	—[b]
Radiologists (Bai et al [24])							
Total	6	0.79 (0.64-0.89)	0.88 (0.78-0.94)	—	—	—	0.85
Junior[c]	3	0.80 (0.72-0.87)	0.88 (0.83-0.92)	—	—	—	—
Senior[d]	3	0.78 (0.70-0.85)	0.87 (0.82-0.91)	—	—	—	—
Junior + AI[e]	—	0.88 (0.81-0.93)	0.93 (0.89-0.96)	—	—	—	—
Senior + AI	—	0.88 (0.81-0.93)	0.89 (0.84-0.93)	—	—	—	—
Radiologists (Zhang et al [33])							
Total	8	0.75 (0.65-0.84)	0.90 (0.86-0.94)	—	—	—	—
Junior	4	0.65 (0.48-0.79)	0.89 (0.81-0.94)	—	—	—	0.82
Senior	4	0.85 (0.70-0.94)	0.91 (0.85-0.96)	—	—	—	0.90
Junior + AI	—	0.80 (0.64-0.90)	0.94 (0.88-0.97)	—	—	—	0.90
Radiologist (Ardakani et al [22]; senior)	1	0.89 (0.81-0.94)	0.83 (0.74-0.89)	—	—	—	—
Other types of pneumonia: deep learning model	7	0.93 (0.92-0.94)	0.95 (0.94-0.95)	22.45 (12.86-39.19)	0.06 (0.03-0.13)	0.98	—
Normal: deep learning model	6	0.95 (0.94-0.96)	0.98 (0.97-0.98)	47.47 (20.70-108.86)	0.04 (0.02-0.08)	0.99	—

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

[b]Not reported.

[c]Junior radiologists have 5 to 15 years of experience.

[d]Senior radiologists have 15 to 25 years of experience.

[e]AI: artificial intelligence.

Figure 2. Performance of the deep learning model for detecting COVID-19.

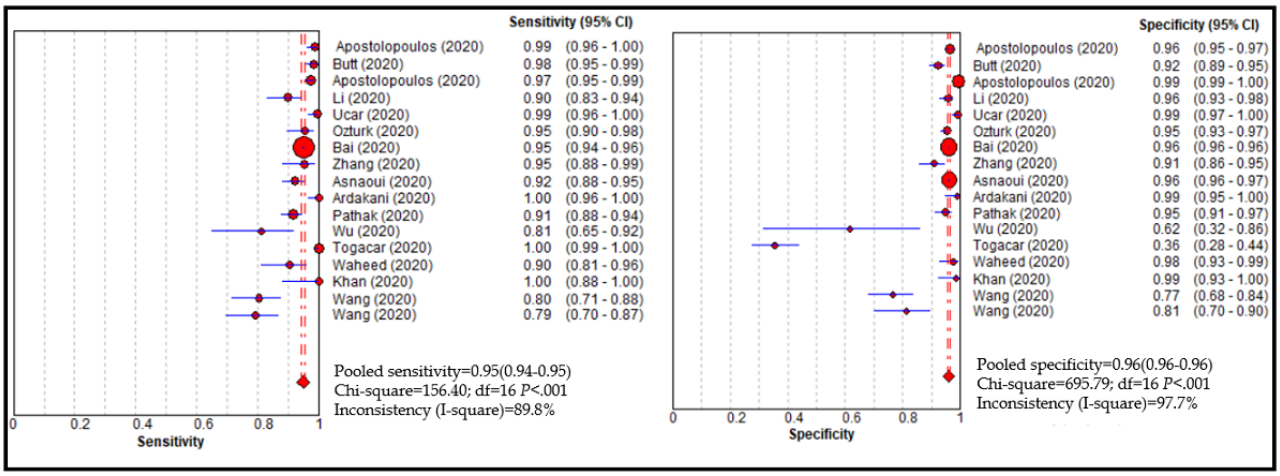
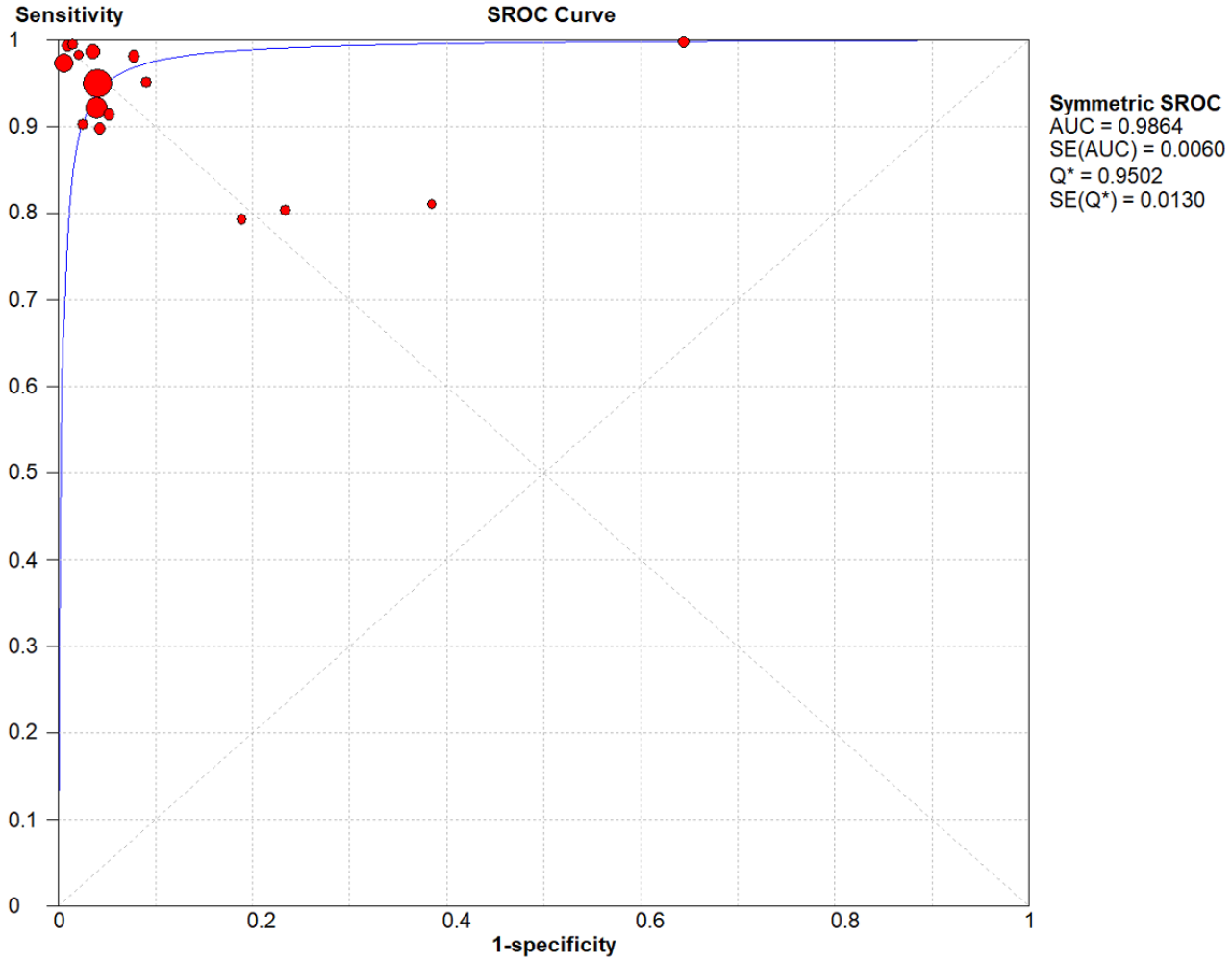


Figure 3. Summary receiver operating characteristic (SROC) curve of the deep learning method. AUC: area under the curve; Q*: this index is defined by the point where sensitivity and specificity are equal.



Performance of Radiologists

Overview

A total of 3 studies compared the performance of DL models with radiologists [22,24,33]. Zhang et al [33] included 8 radiologists with 5 to 25 years of experience; they were categorized into two groups: junior radiologists had 5 to 15 years of experience and senior radiologists had 15 to 25 years of experience. Bai et al [24] compared DL model performance with 6 radiologists; 3 of them had 10 years of experience (ie, junior) and 3 had 20 years of experience (ie, senior). Finally, Ardakani et al [22] compared the performance of DL models with 1 senior radiologist, who had 15 years of experience. The performance of 15 radiologists in detecting COVID-19 was evaluated; the pooled sensitivity and specificity for detecting COVID-19 ranged from 0.75 to 0.89 and from 0.83 to 0.90, respectively. With the assistance of DL-based artificial intelligence (AI) tools, the performance of the junior radiologists improved: sensitivity improved by 0.08 to 0.15 and specificity improved by 0.05.

Sensitivity Analysis

A total of 8 studies evaluated the performance of DL algorithms for detecting COVID-19 using x-ray photographs. The pooled sensitivity and specificity of DL algorithms for detecting COVID-19 were 0.96 (95% CI 0.95-0.97) and 0.97 (95% CI 0.97-0.98), respectively, with an SROC of 0.99. Moreover, 8 studies assessed the performance of DL algorithms for classifying COVID-19 using CT images. The pooled sensitivity and specificity were 0.94 (95% CI 0.94-0.95) and 0.95 (95% CI 0.95-0.96), respectively, with an SROC of 0.96 (see Figures S1-S12 in Multimedia Appendix 1).

Risk of Bias and Applicability

In this meta-analysis, we also assessed heterogeneous findings that originated from included studies based on the QUADAS-2 tool (see Table 3 [13,14,20-33]). The risk of bias for patient selection was unclear for 16 studies. All studies had an unclear risk of bias for flow and timing and for the index test. Moreover, all studies had a high risk of bias for the reference standard. In the case of applicability, all studies had a low risk of bias for patient selection. However, the risk of index test and the applicability concern for the reference standard were uncertain.

Table 3. Quality Assessment of Diagnostic Accuracy Studies-2 for included studies.

Study	Risk of bias (high, low, or unclear)				Applicability concerns		
	Patient selection	Index test	Reference standard	Flow and timing	Patient selection	Index test	Reference standard
Apostolopoulos and Mpesiana [14]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Butt et al [13]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Apostolopoulos et al [21]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Li et al [25]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Ucar and Korkmaz [29]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Ozturk et al [26]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Bai et al [24]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Zhang at al [33]	High	Unclear	High	Unclear	Low	Unclear	Unclear
El Asnaoui and Chawki [20]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Ardakani et al [22]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Pathak et al [27]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Wu et al [32]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Toğaçar et al [28]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Waheed at al [30]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Khan et al [23]	High	Unclear	High	Unclear	Low	Unclear	Unclear
Wang et al [31]	High	Unclear	High	Unclear	Low	Unclear	Unclear

Discussion

Principal Findings

In this study, we evaluated the performance of the DL model regarding detection of COVID-19 automatically using chest images to assist with proper diagnosis and prognosis. The findings of our study showed that the DL model achieved high sensitivity and specificity (95% and 96%, respectively) when detecting COVID-19. The pooled SROC value of both COVID-19 and other types of pneumonia was 98%. The performance of the DL model was comparable to that of experienced radiologists, whose clinical experience was at least 10 years, and the model could improve the performance of junior radiologists.

Clinical Implications

The rate of COVID-19 cases has been mounting day by day; therefore, it is important to quickly and accurately diagnose patients so that we may combat this pandemic. However, screening an increased number of chest images is challenging for the radiologists, and the number of trained radiologists is not sufficient, especially in underdeveloped and developing countries [34]. The recent success of DL applications in imaging analysis of CT scans, as well as x-ray imaging in automatic segmentation and classification in the radiology domain, has encouraged health care providers and researchers to exploit the advancement of deep neural networks in other applications [35]. DL models have been trained to assist radiologists in achieving higher interrater reliability during their years of experience in clinical practice.

Since the start of the COVID-19 pandemic, efforts have been made by AI researchers and AI modelers to help radiologists in the rapid diagnosis of COVID-19 in order to combat the COVID-19 pandemic [33,36]. Developing an accurate, automated AI COVID-19 detection tool is deemed as essential in reducing unnecessary waiting times, shortening screening and examination times, and improving performance. Moreover, such a tool could help to reduce radiologists' workloads and allow them to respond to emergency situations rapidly and in a cost-effective manner [25]. RT-PCR is considered the gold standard detection method; however, findings of our study showed that chest CT could be used as a reliable and rapid approach for screening of COVID-19. Our findings also showed that the DL model was able to discriminate COVID-19 from other types of pneumonia with high a sensitivity and specificity, which is a challenging task for radiologists [32].

Strengths and Limitations

Our study has several strengths. First, this is the first meta-analysis that evaluated the performance of a DL model to classify COVID-19 patients. Second, we considered only peer-reviewed articles to be included in our study because articles that are not peer reviewed might contain bias. Third, we compared the performance of the DL model with that of senior and junior radiologists, which would be helpful for policy makers in considering an automated classification system in real-world clinical settings in order to speed up routine examination.

However, our study also has some limitations that need to be addressed. First, only 16 studies were used to evaluate the performance of the model; inclusion of more studies may have

provided more specific findings. Second, some studies included similar data sets, which may have created some bias, but the researchers in those studies had optimized algorithms to improve performance. Third, two different kinds of digital photographs (ie, CT scan and x-ray) were used to develop and evaluate the performance of the DL model in classifying COVID-19; however, the performance of the DL model was almost the same in both cases. Finally, none of the studies included external validation; therefore, model performance could vary if those models were implemented in other clinical settings.

Future Perspective

The primary objectives of prediction models are the quick screening of COVID-19 patients and to help physicians make appropriate decisions. Misdiagnosis could have a destructive effect on society, as COVID-19 could spread from infected people to healthy people. Therefore, it is important to select a target population among which this automated tool could serve a clinical need; it is also important to select a representative data set on which the model could be trained, developed, and validated internally and externally. All the studies included in this meta-analysis had a high risk of bias for patient selection and reference standards. Moreover, generalizability was lacking in the newly developed classification models. Models without proper evidence and with a lack of external validation are not appropriate for clinical practice because they might cause more harm than good. Since the number of cases is mounting each day and COVID-19 is spreading to all continents, it is therefore important to develop a model to assist in the quick and efficient screening of patients during the COVID-19 pandemic. This could encourage clinicians and policy makers to prematurely implement prediction models without sufficient documentation and validation. All studies showed promising discrimination in their training, testing, and validation cohorts, but future studies should focus on external validation and comparing their findings to other data sets. Interpretability of DL systems is more important to a health care professional than to an AI expert. Proper interpretation and explanation of algorithms will more likely be acceptable to physicians. More clinical research is needed to determine the tangible benefits for patients in terms of the high performance of the model. High sensitivity and specificity do not necessarily represent clinical efficacy, and the higher value of the AUROC is not always the best metric to exhibit clinical applicability. All papers should follow standard guidelines and they should present positive and negative predictive values in order to be able to make a fair comparison. Although all of the included studies used a significant amount of data to compare model performance to that of the radiologists, they used only retrospective data to train the models, which might result in worse performance in real-world clinical settings, as data complexity is different. Therefore, prospective evaluation is needed in future studies before considering implementation in clinical settings. AI models always consist of potential flaws, including the inapplicability of new data, reliability, and bias. Generalization of the model is important for presenting the real performance because the rate of sensitivity and specificity varied across the studies (0.79 to 1.00 and 0.62 to 1.00, respectively). A higher number of false negatives will make the situation worse and will waste health care resources.

Conclusions

Our study showed that the DL model had immense potential to distinguish COVID-19 patients, with high sensitivity and specificity, from patients with other types of pneumonia and normal patients. DL-based tools could assist radiologists in the fast screening of COVID-19 and in classifying potential high-risk patients, which could have clinical significance for the early management of patients and could optimize medical resources. A higher number of false negatives could have a devastating effect on society; therefore, it is crucial to test the performance of models with other, unknown data sets. Retrospective evaluation and reliable interpretation are warranted to consider the application of AI models in real-world clinical settings.

Acknowledgments

This research is sponsored, in part, by the Ministry of Education (MOE) (grants MOE 108-6604-001-400 and DP2-109-21121-01-A-01) and the Ministry of Science and Technology (MOST) (grants MOST 108-2823-8-038-002- and 109-2222-E-038-002-MY2).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary figures.
[DOCX File , 388 KB - medinform_v9i4e21394_app1.docx]

References

1. Sun J, He W, Wang L, Lai A, Ji X, Zhai X, et al. COVID-19: Epidemiology, evolution, and cross-disciplinary perspectives. Trends Mol Med 2020 May;26(5):483-495 [FREE Full text] [doi: 10.1016/j.molmed.2020.02.008] [Medline: 32359479]
2. Park M, Cook AR, Lim JT, Sun Y, Dickens BL. A systematic review of COVID-19 epidemiology based on current evidence. J Clin Med 2020 Mar 31;9(4):967 [FREE Full text] [doi: 10.3390/jcm9040967] [Medline: 32244365]

3. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) — China, 2020. China CDC Wkly 2020 Feb 14;2(8):113-122 [FREE Full text] [doi: 10.46234/ccdcw2020.032]
4. Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun 2020 May;109:102433 [FREE Full text] [doi: 10.1016/j.jaut.2020.102433] [Medline: 32113704]
5. 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]
6. Kojima N, Turner F, Slepnev V, Bacelar A, Deming L, Kodeboyina S, et al. Self-collected oral fluid and nasal swab specimens demonstrate comparable sensitivity to clinician-collected nasopharyngeal swab specimens for the detection of SARS-CoV-2. Clin Infect Dis 2020 Oct 19:ciaa1589 [FREE Full text] [doi: 10.1093/cid/ciaa1589] [Medline: 33075138]
7. Rhoads DD, Cherian SS, Roman K, Stempak LM, Schmotzer CL, Sadri N. Comparison of Abbott ID Now, DiaSorin Simplexa, and CDC FDA emergency use authorization methods for the detection of SARS-CoV-2 from nasopharyngeal and nasal swabs from individuals diagnosed with COVID-19. J Clin Microbiol 2020 Jul 23;58(8):1-2. [doi: 10.1128/jcm.00760-20]
8. Jones BP, Tay ET, Elikashvili I, Sanders JE, Paul AZ, Nelson BP, et al. Feasibility and safety of substituting lung ultrasonography for chest radiography when diagnosing pneumonia in children: A randomized controlled trial. Chest 2016 Jul;150(1):131-138 [FREE Full text] [doi: 10.1016/j.chest.2016.02.643] [Medline: 26923626]
9. Ye X, Xiao H, Chen B, Zhang S. Accuracy of lung ultrasonography versus chest radiography for the diagnosis of adult community-acquired pneumonia: Review of the literature and meta-analysis. PLoS One 2015;10(6):e0130066 [FREE Full text] [doi: 10.1371/journal.pone.0130066] [Medline: 26107512]
10. Bai HX, Hsieh B, Xiong Z, Halsey K, Choi JW, Tran TML, et al. Performance of radiologists in differentiating COVID-19 from non-COVID-19 viral pneumonia at chest CT. Radiology 2020 Aug;296(2):E46-E54 [FREE Full text] [doi: 10.1148/radiol.2020200823] [Medline: 32155105]
11. Ai T, Yang Z, Hou H, Zhan C, Chen C, Lv W, et al. Correlation of chest CT and RT-PCR testing for coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. Radiology 2020 Aug;296(2):E32-E40 [FREE Full text] [doi: 10.1148/radiol.2020200642] [Medline: 32101510]
12. Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: A descriptive study. Lancet Infect Dis 2020 Apr;20(4):425-434. [doi: 10.1016/s1473-3099(20)30086-4]
13. Notice of retraction: Butt C, Gill J, Chun D, Babu BA. Deep learning system to screen coronavirus disease 2019 pneumonia. Appl Intell 2020 Apr 22:1-7. [doi: 10.1007/s10489-020-01714-3]
14. Apostolopoulos ID, Mpesiana TA. Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med 2020 Jun;43(2):635-640 [FREE Full text] [doi: 10.1007/s13246-020-00865-4] [Medline: 32524445]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med 2009 Jul 21;6(7):e1000097. [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
16. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, QUADAS-2 Group. QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011 Oct 18;155(8):529-536. [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
17. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytic approaches and some additional considerations. Stat Med 1993 Jul 30;12(14):1293-1316. [doi: 10.1002/sim.4780121403] [Medline: 8210827]
18. Islam MM, Yang H, Poly TN, Jian W, Jack Li YC. Deep learning algorithms for detection of diabetic retinopathy in retinal fundus photographs: A systematic review and meta-analysis. Comput Methods Programs Biomed 2020 Jul;191:105320. [doi: 10.1016/j.cmpb.2020.105320] [Medline: 32088490]
19. Walter SD. Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat Med 2002 May 15;21(9):1237-1256. [doi: 10.1002/sim.1099] [Medline: 12111876]
20. El Asnaoui K, Chawki Y. Using x-ray images and deep learning for automated detection of coronavirus disease. J Biomol Struct Dyn 2020 May 22:1-12 [FREE Full text] [doi: 10.1080/07391102.2020.1767212] [Medline: 32397844]
21. Apostolopoulos ID, Aznaouridis SI, Tzani MA. Extracting possibly representative COVID-19 biomarkers from x-ray images with deep learning approach and image data related to pulmonary diseases. J Med Biol Eng 2020 May 14;40:1-8 [FREE Full text] [doi: 10.1007/s40846-020-00529-4] [Medline: 32412551]
22. Ardakani AA, Kanafi AR, Acharya UR, Khadem N, Mohammadi A. Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks. Comput Biol Med 2020 Jun;121:103795 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103795] [Medline: 32568676]
23. Khan AI, Shah JL, Bhat MM. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput Methods Programs Biomed 2020 Nov;196:105581 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105581] [Medline: 32534344]

24. Bai HX, Wang R, Xiong Z, Hsieh B, Chang K, Halsey K, et al. Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT. Radiology 2020 Sep;296(3):E156-E165 [FREE Full text] [doi: 10.1148/radiol.2020201491] [Medline: 32339081]
25. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, et al. Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: Evaluation of the diagnostic accuracy. Radiology 2020 Aug;296(2):E65-E71 [FREE Full text] [doi: 10.1148/radiol.2020200905] [Medline: 32191588]
26. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O, Acharya UR. Automated detection of COVID-19 cases using deep neural networks with x-ray images. Comput Biol Med 2020 Jun;121:103792 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103792] [Medline: 32568675]
27. Pathak Y, Shukla P, Tiwari A, Stalin S, Singh S, Shukla P. Deep transfer learning based classification model for COVID-19 disease. Ing Rech Biomed 2020 May 20:1-7 [FREE Full text] [doi: 10.1016/j.irbm.2020.05.003] [Medline: 32837678]
28. Toğaçar M, Ergen B, Cömert Z. COVID-19 detection using deep learning models to exploit social mimic optimization and structured chest x-ray images using fuzzy color and stacking approaches. Comput Biol Med 2020 Jun;121:103805 [FREE Full text] [doi: 10.1016/j.compbiomed.2020.103805] [Medline: 32568679]
29. Ucar F, Korkmaz D. COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from x-ray images. Med Hypotheses 2020 Jul;140:109761 [FREE Full text] [doi: 10.1016/j.mehy.2020.109761] [Medline: 32344309]
30. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR. CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. IEEE Access 2020;8:91916-91923. [doi: 10.1109/access.2020.2994762]
31. Wang S, Zha Y, Li W, Wu Q, Li X, Niu M, et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. Eur Respir J 2020 Aug;56(2):2000775 [FREE Full text] [doi: 10.1183/13993003.00775-2020] [Medline: 32444412]
32. Wu X, Hui H, Niu M, Li L, Wang L, He B, et al. Deep learning-based multi-view fusion model for screening 2019 novel coronavirus pneumonia: A multicentre study. Eur J Radiol 2020 Jul;128:109041 [FREE Full text] [doi: 10.1016/j.ejrad.2020.109041] [Medline: 32408222]
33. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography. Cell 2020 Sep 03;182(5):1360 [FREE Full text] [doi: 10.1016/j.cell.2020.08.029] [Medline: 32888496]
34. Liedenbaum MH, Bipat S, Bossuyt PMM, Dwarkasing RS, de Haan MC, Jansen RJ, et al. Evaluation of a standardized CT colonography training program for novice readers. Radiology 2011 Feb;258(2):477-487. [doi: 10.1148/radiol.10100019] [Medline: 21177395]
35. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-9389. [doi: 10.1109/access.2017.2788044]
36. Hemdan E, Shouman M. COVIDX-Net: A framework of deep learning classifiers to diagnose COVID-19 in x-ray images. arXiv. Preprint posted online on March 24, 2020. [FREE Full text]

Abbreviations

AI: artificial intelligence
AUC: area under the curve
AUROC: area under the receiver operating characteristic curve
CT: computed tomography
DL: deep learning
DOR: diagnostic odds ratio
MOE: Ministry of Education
MOST: Ministry of Science and Technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies-2
ROC: receiver operating characteristic
RT-PCR: reverse transcription–polymerase chain reaction
SROC: summary receiver operating characteristic

Edited by C Lovis; submitted 16.06.20; peer-reviewed by CT Cheng, E Frontoni; comments to author 26.08.20; revised version received 04.09.20; accepted 21.03.21; published 29.04.21.

Please cite as:
Poly TN, Islam MM, Li YCJ, Alsinglawi B, Hsu MH, Jian WS, Yang HC
Application of Artificial Intelligence for Screening COVID-19 Patients Using Digital Images: Meta-analysis
JMIR Med Inform 2021;9(4):e21394
URL: https://medinform.jmir.org/2021/4/e21394
doi:10.2196/21394
PMID:33764884

Original Paper

Returning to a Normal Life via COVID-19 Vaccines in the United States: A Large-scale Agent-Based Simulation Study

Junjiang Li[1]; Philippe Giabbanelli[1], BSc, MSc, PhD

Department of Computer Science & Software Engineering, Miami University, Oxford, OH, United States

Corresponding Author:
Philippe Giabbanelli, BSc, MSc, PhD
Department of Computer Science & Software Engineering
Miami University
205 Benton Hall
Oxford, OH, 45056
United States
Phone: 1 513 529 0147
Email: aqualonne@free.fr

Abstract

Background: In 2020, COVID-19 has claimed more than 300,000 deaths in the United States alone. Although nonpharmaceutical interventions were implemented by federal and state governments in the United States, these efforts have failed to contain the virus. Following the Food and Drug Administration's approval of two COVID-19 vaccines, however, the hope for the return to normalcy has been renewed. This hope rests on an unprecedented nationwide vaccine campaign, which faces many logistical challenges and is also contingent on several factors whose values are currently unknown.

Objective: We study the effectiveness of a nationwide vaccine campaign in response to different vaccine efficacies, the willingness of the population to be vaccinated, and the daily vaccine capacity under two different federal plans. To characterize the possible outcomes most accurately, we also account for the interactions between nonpharmaceutical interventions and vaccines through 6 scenarios that capture a range of possible impacts from nonpharmaceutical interventions.

Methods: We used large-scale, cloud-based, agent-based simulations by implementing the vaccination campaign using COVASIM, an open-source agent-based model for COVID-19 that has been used in several peer-reviewed studies and accounts for individual heterogeneity and a multiplicity of contact networks. Several modifications to the parameters and simulation logic were made to better align the model with current evidence. We chose 6 nonpharmaceutical intervention scenarios and applied the vaccination intervention following both the plan proposed by Operation Warp Speed (former Trump administration) and the plan of one million vaccines per day, proposed by the Biden administration. We accounted for unknowns in vaccine efficacies and levels of population compliance by varying both parameters. For each experiment, the cumulative infection growth was fitted to a logistic growth model, and the carrying capacities and the growth rates were recorded.

Results: For both vaccination plans and all nonpharmaceutical intervention scenarios, the presence of the vaccine intervention considerably lowers the total number of infections when life returns to normal, even when the population compliance to vaccines is as low as 20%. We noted an unintended consequence; given the vaccine availability estimates under both federal plans and the focus on vaccinating individuals by age categories, a significant reduction in nonpharmaceutical interventions results in a counterintuitive situation in which higher vaccine compliance then leads to more total infections.

Conclusions: Although potent, vaccines alone cannot effectively end the pandemic given the current availability estimates and the adopted vaccination strategy. Nonpharmaceutical interventions need to continue and be enforced to ensure high compliance so that the rate of immunity established by vaccination outpaces that induced by infections.

(JMIR Med Inform 2021;9(4):e27419) doi:10.2196/27419

KEYWORDS

agent-based model; cloud-based simulations; COVID-19; large-scale simulations; vaccine; model; simulation; United States; agent-based; effective; willingness; capacity; plan; strategy; outcome; interaction; intervention; scenario; impact

Introduction

The Centers for Disease Control and Prevention (CDC) forecasted that 300,000 deaths would be attributable to COVID-19 by the end of the year. Reality defied expectations, as COVID-19 was *directly* responsible for approximately 350,000 deaths in the United States out of 20 million *reported* cases (for forecasts and total case numbers, see [1]), which may only represent one out of seven actual cases based on CDC estimates for September 2020 [2]. Despite popular comparison with the flu, the ongoing COVID-19 epidemic has thus already claimed five times as many lives than the worst year for the flu, whose recent yearly death tolls range from a low of 16,000 to a high of 68,000 [3]. To contextualize the impact of COVID-19, we noted that the US *life expectancy* decreased by more than a year, which is ten times worse than the decline from the opioid epidemic [4]. In another comparison, 2020 is the *largest single-year increase in mortality* in the United States since 1918, which had both a flu pandemic and a war. This reflects both direct and *indirect* consequences of COVID-19, such as disrupting in-person treatments [5] and supply networks, with effects as far ranging as an increase in drug overdose [6]. To complement measures of short-term effects such as deaths or number of cases, we also noted the long-term impacts captured by the outpatient journey. Common symptoms often persist over a month (eg, fatigue, cough, headache, sore throat, or loss of smell) [7-9], and less frequent ones can be severe since COVID-19 involves many organs. Effects can involve the cardiovascular system in up to 20%-30% of patients who are hospitalized [10,11] (eg, cardiac injury, vascular dysfunction, or thrombosis), result in kidney injury [10] or pulmonary abnormalities [12], or lead to a deterioration in cognition due to cerebral microstructural changes [13]. Based on similar infections, such effects can be long: for instance, inflammation of the heart caused by viral infections (eg, myocarditis) can have a recovery period spanning months to years.

Interventions in 2020 were strictly *nonpharmaceutical*, as vaccines were being developed and tested. Such intervention strategies have included preventative care (eg, social distancing, handwashing, and face masks), lockdowns (eg, travel restrictions, school closures, and remote work), and logistics associated with testing (eg, contact tracing and quarantine) [14,15]. The range of nonpharmaceutical interventions adopted at various times across countries can be seen in further details through the CoronaNet project [16] or the collection of essays "mobilizing policy (in)capacity to fight COVID-19" published in mid-2020 [17]. In early 2021, two vaccines were deployed (Pfizer-BioNTech and Moderna) with plans for up to three additional vaccines (AstraZeneca, Janssen, and Novavax) [18]. With the availability of vaccines comes the key question: when will life return to normal in the United States? The implicit expectation is to see a return to normalcy thanks to the vaccine, rather than due to a high number of cases with its accompanying death toll.

In a highly publicized interview, Dr Anthony Fauci, director of the National Institute of Allergy and Infectious Diseases, estimated a return to normal by fall, *if the vaccination campaign is successful* [19]. Getting a precise estimate of when life will return to normal is a challenge, as it depends on numerous interrelated factors: potential behavioral changes affecting nonpharmaceutical approaches (eg, lesser compliance to mask wearing and social distancing), participation in the vaccination campaign, logistics associated with vaccination (ie, who can get vaccinated and when), and mutations leading to new strains with different biological properties (eg, higher infectivity) or unknown vaccine responses. In this paper, we use large-scale simulations to identify *when* there will be an inflection point in the dynamics of the disease and the *level* of cases that will be obtained.

Simulations have been used since the early days of the COVID-19 pandemic. Classic compartmental epidemiological models were first produced (eg, many susceptible-exposed-infected-removed models [20-23]), with a focus on estimating broad trends and key epidemiological quantities such as the expected number of new cases generated by each infected individual (ie, the basic reproduction number R_0). Such compartmental models provide limited support to study the effect of interventions, for instance by lowering the contact rate to represent the impact of social distancing. A research shift in the second part of 2020 resulted in the growing use of *agent-based models (ABMs)* to support the analysis of interventions by explicitly modeling each individual and their interactions among each other or with the environment. This shift to individual-level models was underpinned by the evidence of *heterogeneity* in risk factors (eg, older age, hypertension, respiratory disease, and cardiovascular disease [24,25]) and behaviors (eg, noncompliance with social distancing orders) based on personal beliefs and values [26,27]. There is also spatial variation in socio-ecological vulnerability to COVID-19 [28], with rural counties being at higher risk (due to eg, older population with more underlying conditions and lower access to resources) [29,30] and hence experiencing higher mortality rates [31]. Finally, there is a documented heterogeneity in transmission based on contact tracing data [32], which stresses the need to use realistic networks when modeling the spread of COVID-19 [33]. Considering this growing evidence base, our study relies on an ABM, which accounts for individual heterogeneity (eg, in age), explicitly embeds them in a network to model their contacts, and simultaneously considers different network types (eg, community and work) to account for various settings.

By adding vaccines to a previously validated ABM of COVID-19, we are able to assess how the number and timing of cases depends on key factors such as the population's interest in vaccines and the efficacy of vaccines. Our specific contributions are twofold:

1. We extend the validated COVASIM model with a detailed process of vaccination, accounting for vaccine efficacy, interest in vaccination, and fluctuations in vaccination capacity. Our process models the need for two doses and the possibility of being infected until the second dose is administered.
2. We examine vaccination interventions under two hypotheses for the number of doses available and considering concurrent nonpharmaceutical interventions.

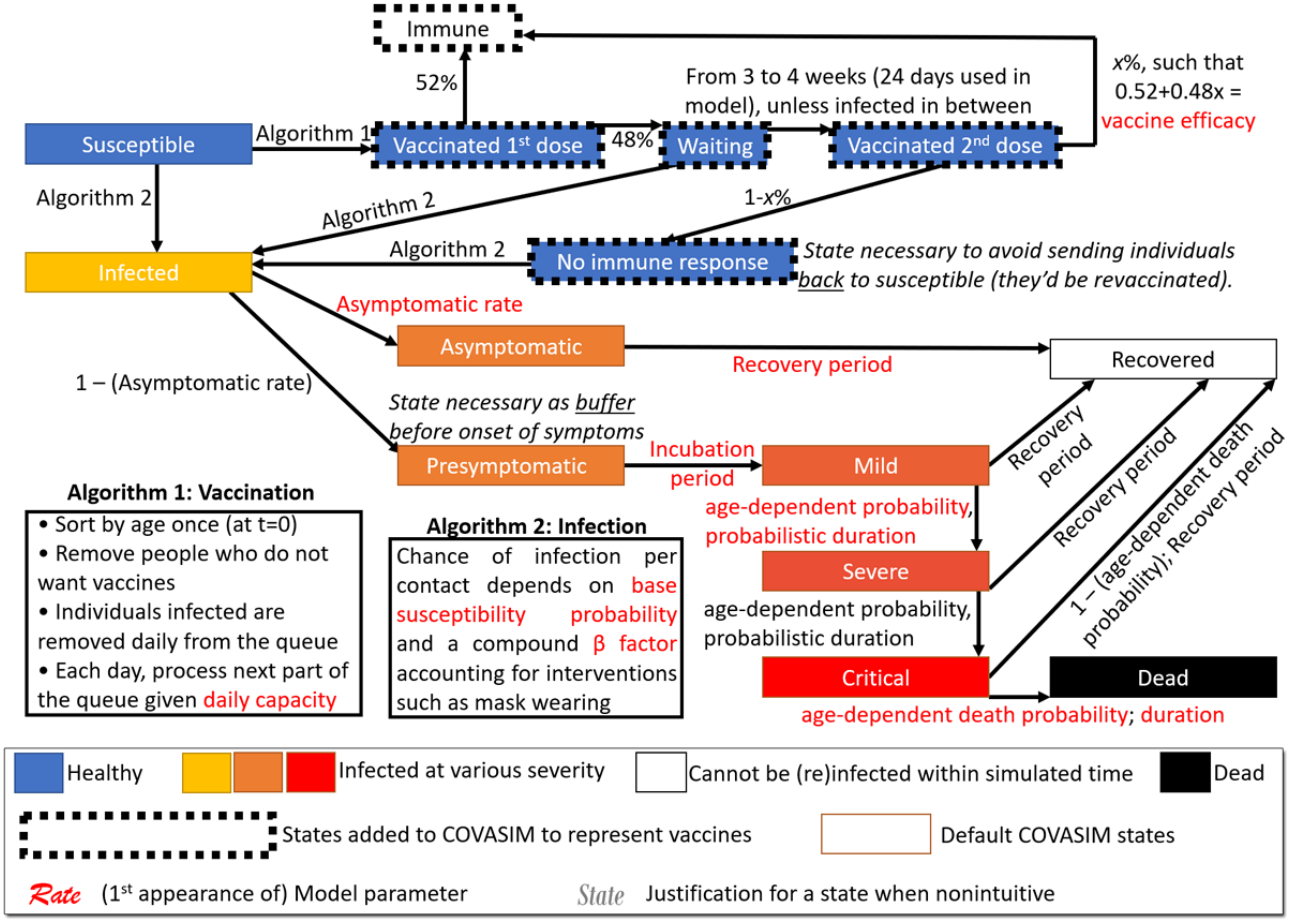
The remainder of this paper is structured as follows. In our methods, we briefly cover the rationale for choosing COVASIM and how we adapted the model to account for the latest epidemiological evidence. We then explain which nonpharmaceutical interventions are simulated, in line with our previous work [34]. Most importantly, we detail the novel extension of vaccines into COVASIM and our examination of the trends in cumulative infections using a logistic growth model. The following section presents and analyzes our results. Our final section discusses our main findings and provides an exhaustive list of limitations due to the ongoing nature of the pandemic and challenges in vaccination.

Methods

Overview

COVASIM was developed under leadership of the Institute for Disease Modeling and released in May 2020 by Kerr and colleagues [35]. It is one of several open-source ABMs, together with OpenABM-Covid19 [36] or COMOKIT [37]. The model captures the transition from susceptible to infected followed by a split between asymptomatic individuals and various degrees of symptoms, resulting either in recovery or death (Figure 1). The model was created to support interventions offered at the time, which did not include vaccination. We thus *modified the model* to account for our current understanding of viral dynamics and the use of vaccines over two doses (Figure 1). When instantiating the model to the US population, we used a resolution of 1:500 (ie, each simulated agent accounts for 500 US inhabitants). Given our resolution and target population size, our application exceeded half a million agents and can thus be described as a "*large-scale COVID-19 simulation*" [38]. Our simulations started on January 1, 2020, using CDC data for the number of infected, recovered, and immunized individuals to date (see subsection Initializing the Model). We then simulate for 6 months, that is, 180 time ticks based on a temporal resolution of 1 day per simulation step (ie, *tick*). To cope with the computational challenges created by a large-scale stochastic model, a philanthropic grant supports us in performing cloud-based simulations via the Microsoft Azure (Microsoft Corporation) platform.

Figure 1. Overview of our modified COVASIM model containing the state diagram and specification of all transitions, including key procedures for vaccination and infection.



The COVASIM Model: Rationale for Selection and Evidence-Based Updates

Apart from being open source, there are two reasons that we selected COVASIM. First, it captures heterogeneity within individuals (eg, assigns an age and uses age-specific disease outcomes) and transmission patterns by placing agents within synthetic networks corresponding to a multiplicity of contexts: work (based on employment rates), school (based on enrollment), home (based on household size), and the general community. However, these high-resolution age-specific contact patterns are not unique to COVASIM. For example, the

OpenABM-Covid19 [36] also embeds agents in age-stratified occupation networks (encompassing work and school), household networks, and a *general* random network. COMOKIT [37] similarly uses the Gen* toolkit from the same team to redistribute populations from census units down to exact buildings such as the nearest school. Thus, the second rationale for choosing this platform is that it has been used in the most peer-reviewed modeling studies to date [39,40], hence providing an additional layer of scrutiny and confidence in the correctness of the model (ie, validation) and its implementation (ie, verification). As detailed in our recent study [34], changes in the evidence base have required alteration in the model to keep it valid. Consequently, we modified three COVASIM parameters to account for the current biological and epidemiological evidence on COVID-19 (Table 1).

Table 1. Adjusted parameters based on reports in the United States.

COVASIM construct	Initial value	Modified value	Rationale for modification
Incubation: delay from infection to viral shedding	Lognormal(4.6, 4.8)	Lognormal(4.1, 4.8)	The combined distribution of the incubation period did not match the latest evidence. The adjustment aligns it with the evidence.
Incubation: delay from viral shedding to onset of symptoms	Lognormal(1,1)	Lognormal(1, 1.8)	Same as above
Proportion of symptomatic cases	0.7	0.6	Although reports vary, Dr Fauci stated that 40% of the US cases were asymptomatic.

Selection and Representation of Concurrent Nonpharmaceutical Interventions

In addition to support for heterogeneity, COVASIM implements several nonpharmaceutical interventions. Although our focus is on vaccines, such interventions may be continuing in parallel with the vaccination campaign; hence, we have to take them into account when forecasting case counts. Interventions can be organized into three broad categories: preventative care (eg, *social distancing* and *face masks*), lockdown (eg, *stay-at-home* orders such as remote work or school closures), or testing-related (eg, *testing* itself, then *quarantining* and *contact tracing*) [14,41,42]. In line with our previous work on nonpharmaceutical interventions, we considered all 6 specific interventions. Although all 6 are natively supported by the COVASIM platform, we changed testing delays from their default value (constant) to a distribution (based on a survey across all 50 US states) [43], thus accounting for the variability observed in practice.

Since our focus is on vaccines, our search space is primarily devoted to quantifying the effect of vaccine-related variables (ie, efficacy, compliance, and capacity). As every nonpharmaceutical intervention could lead to several variables (eg, compliance with face masks or efficacy of face masks), considering all variables for every such intervention *in addition to* vaccine-related variables would lead to an impractical search space. We thus leveraged the systematic assessment of our previous study [34], which simulated all combinations of nonpharmaceutical interventions at two different levels of strength (ie, a binary factorial design of experiments). We analyzed results from this broad search to select 5 scenarios (Table 2) that resulted in five different levels of infection count after 6 months, in the absence of any vaccine (Figure 2). In other words, to circumvent the unwieldy notion of simulating all aspects of vaccines and nonpharmaceutical interventions, we selected 5 scenarios that produce linear to logistic growths in cumulative infections, thereby conducting a parameter sweep across possible growth behaviors. We supplemented these 5 scenarios with an extreme *no intervention* scenario, which provides an upper bound on the number of cases.

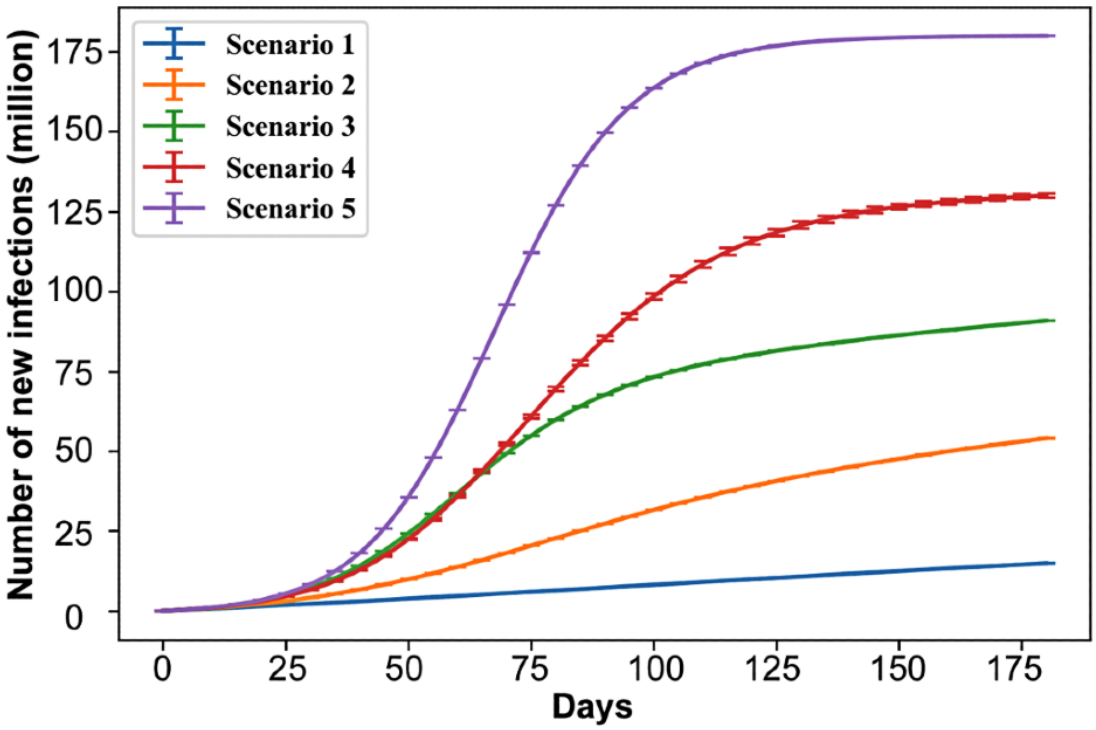
Table 2. Scenarios depicting concurrent nonpharmaceutical interventions, chosen for their ability to create five markedly different outcomes together with a nonintervention case.

Features	Scenario					
	1	2	3	4	5	6 (do nothing)
Networks impacted	Work, school	Work, school	Community	Community	Community	All
Contact in work and school (as a function of default; %)	70	95	N/A[a]	N/A	N/A	100
Contact in community (as a function of default; %)	N/A	N/A	70	70	90	100
Daily tests[b]	1,110,000	600,000	600,000	1,110,000	600,000	No testing
A positive test leads to quarantine. Is a second test required to end quarantine?	No	Yes	No	Yes	Yes	No testing
Test sensitivity	1	1	1	0.55	0.55	No testing
Ratio of contacts that can be traced	0.2	1	1	0.2	0.2	No tracing
After how many days will contact tracing results arrive (ie, contact tracing delay)?	0	7	7	7	7	No tracing
Starting contact tracing if one has just been tested and exposed (one infected peer)	Yes	No	No	Yes	No	No tracing

[a]N/A: not applicable.

[b]These numbers reflect the total daily capacity at the scale of the US population. As our simulation uses a scale of 1:500, the capacity in the model is scaled down accordingly.

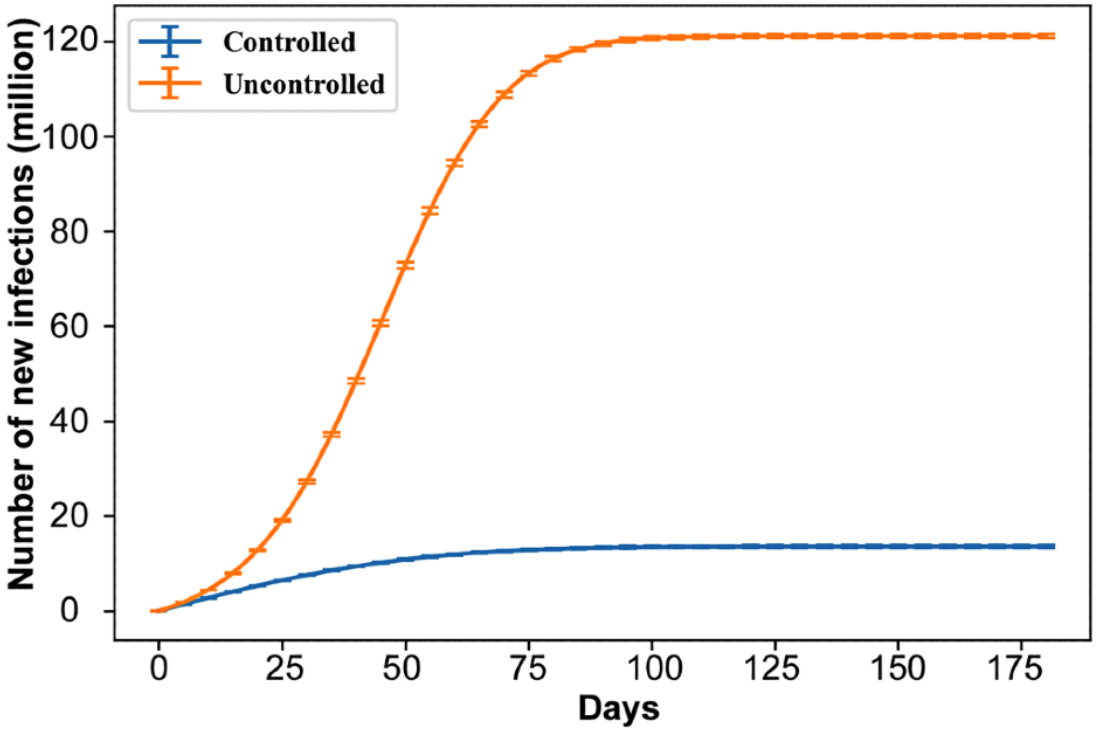
Figure 2. Number of new infections during the simulation (ie, cumulative cases) under five scenarios (each based on a combination of interventions), which were selected for their ability to represent different trends in the number of cases over time, without a vaccine.



Given that we made minor changes to the biology (incubation and proportion of symptomatic cases) and consider several ongoing intervention scenarios, it is necessary to confirm the validity of the model established using earlier data in previously published studies. Consequently, we ran the modified COVASIM model based on data observed until September 3, 2020, and compared the simulated results with observations until the end of year. Similar trends and orders of magnitude were observed (Figure 3), thus providing qualitative validation. Note that the 5 scenarios chosen (Table 2) bound the growth of

COVID-19 in the United States such that we are comprehensively examining possible trends going forward instead of limiting ourselves to the single trend that fit best on previous data.

Figure 3. Comparison of changes in cumulative infections between a COVASIM simulation and reality from September 3, 2020, to the end of 2020. The simulation included a reduction on work and school contacts (set to 95% of their capacity), 600,000 daily and highly sensitive tests, quarantining upon testing, immediate tracing to identify all contacts, and a presumptive approach.



Extending COVASIM With Pharmaceutical Interventions: A Two-step Vaccination

As detailed in our discussion, there is substantial uncertainty and frequent changes regarding the number of vaccines that *may* be administered monthly. We thus considered two vaccine availability scenarios, both proposed by federal governments. The first scenario from the former Trump administration, named Operation Warp Speed, stated that vaccines will be available in tiered amounts (20 million in December, 30 million in January, and 50 million every month thereafter). The second scenario from the Biden administration, known as the *100-day goal*, proposes that there will be 1 million vaccines every day [44], thus covering 50 million Americans. Although there are other scenarios, they vary from state to state (eg, the governor of New Jersey aspires to vaccinate 70% of the adult population within 6 months [45]) and are subject to frequent revisions. Given the countrywide nature of our simulation, we relied on federal plans while detailing challenges (see also the Discussion section).

In setting the monthly capacity, we noticed the necessity to *adjust the schedule* of the Operation Warp Speed plan, since the initial aim of 20 million people immunized by the end of December 2020 only resulted in 3 million doses administered. In other words, it would be incorrect to model the monthly capacity of Operation Warp Speed as announced since there is evidence that its initial objective was unmet, due to a variety of logistical challenges. Consequently, we shifted the expectations of the Operation Warp Speed plan by 1 month, such that the capacity for January now corresponds to the initial expectations for December (20 million) and so forth.

At the same time as either vaccination schedule is active, we also have the 6 scenarios listed in the previous sections. As these scenarios include a no-intervention case, we are able to study the interaction between nonpharmaceutical interventions and vaccines. In total, this gives 12 distinct situations. In addition, we also varied two essential parameters regarding vaccines: the percentage of the population that seeks vaccination (which we refer to as *vaccine compliance* from hereon) and the efficacy of the vaccine. Varying these two parameters across 12 situations in a large-scale ABM results in substantial computing needs. These are challenging to parallelize, as the run time of each experiment is not the same. Therefore, we took advantage of the massive parallelism enabled by the cloud computing platform Azure to accelerate computation. Using this platform, we varied vaccine compliance and vaccine efficacy between the bounds listed in Table 3.

Table 3. Vaccine parameters used in the study. Intermediate values in the interval bounded by the low and high values are automatically explored.

Parameters	Low value (%)	High value (%)
Vaccine compliance	20	60
Vaccine efficacy	88	99

Regarding our approach to vaccine efficacy, we noted that individuals can be infected after their first dose, as has been documented on thousands of cases [46]. We thus used the probability of 52% (observed in clinical trials [47]) to obtain early protection by the vaccine, and otherwise, an individual may still be infected in the waiting period leading to the second dose. After the second dose is applied, we needed to ensure that the agent meets the vaccine efficacy set by our parameters. That is, the probability of obtaining immunity after the second dose was set such that the probability of immunity *from the two doses* matches the vaccine efficacy.

Although we did not track which of the two approved mRNA COVID-19 vaccines (Pfizer-BioNTech or Moderna) were administered, we varied vaccine efficacy to account for a margin of uncertainty regarding their respective performances. Since the vaccine capacity is either planned to increase (Operation Warp Speed) or be at a high constant rate, a simulated agent given one dose will always be able to come back to get the second dose on time. Should an agent be contaminated or die before the second dose, it is then released for administration to another agent.

We also varied the percentage of the population who seeks vaccination. As noted in a recent study, this percentage has varied among studies: 10.8% did not intend to be vaccinated when asked in April 2020, but this number jumped to 31.1% by May, and an August poll found that only a *minority* would want to be vaccinated [48]. In addition to changes in the sociopolitical climate and public discourse surrounding vaccination, there will also be changes since "many receptive participants preferred to wait until others have taken the vaccine" [49]. Seeing positive vaccination outcomes in others may in part address the fear of serious side effects, which is a recurring concern for individuals who may not intend to participate in vaccination [50]. Given past variations and changes in the future, we handled uncertainty through a parameter sweep in vaccine compliance.

Initializing the Model

A simulation model is composed of an initialization (setting characteristics of agents for t=0) and rules governing its update, thereby producing data for analysis. The previous subsections covered the rationale for the inclusion of agents' characteristics and the design of the rules, while the next subsection focuses on the analysis. This subsection thus briefly covers our approach to initialization such that our results could be independently replicated by other modeling teams.

Our initial time tick t=0 corresponds to January 1, 2020. We thus needed to set the number of agents who have been infected, recovered, or immunized (due to the rollout of vaccines in December) by that time. A COVID-19 case remains infectious within a time window of 2 weeks, after which there is either recovery or complications. From December 18-31, there was a total of 3,311,345 active cases. To appropriately initialize our simulation, we needed to further track *when* an individual was infected. Incorrectly setting them to be all infected on December 18 would result in nobody being infected when the simulation starts on January 1. At the other extreme, assuming that they were all infected on December 31 would lead to an overestimate of disease spread into 2021. We thus seeded the timing of each infection by using the daily distribution from CDC data between December 18-31 (Table 4). All numbers were divided by 500 since our agent resolution is 1 agent for 500 real-world US inhabitants (1:500). The number of individuals who acquired immunity via recovery was set to the total case count observed by December 17. Individuals who died from COVID-19 are grouped together with recovered ones (ie, we did not subtract them from the count) since our simulations track the number of new infections; dead individuals do not alter these results as they can neither be infected nor infect others. The total number of individuals immunized from vaccination was set to 2 million (ie, 4000 agents).

Table 4. Timing of the infection in the 2 weeks preceding the start of our simulation, such that our agents can be initialized at the appropriate state of their infection.

Specific day of the infection	Individuals infected, n
December 18, 2020	236,063
December 19, 2020	202,050
December 20, 2020	198,129
December 21, 2020	184,632
December 22, 2020	196,516
December 23, 2020	229,746
December 24, 2020	193,277
December 25, 2020	139,152
December 26, 2020	179,707
December 27, 2020	146,593
December 28, 2020	177,814
December 29, 2020	201,428
December 30, 2020	230,982
December 31, 2020	229,634

Analyzing the Progression of Cumulative Infections Through a Logistic Growth Model

To quantify the spread of the disease, we fitted the progression of cumulative infection to a logistic growth model, which is a simple yet effective model describing resource-limited growths in natural processes and has been used on several occasions for COVID-19 [51-53]. Let the cumulative infection be $P = P(t)$, then the logistic model stipulates that P is the solution of the differential equation:

where  is the time derivative of P, r is the *growth rate* (proportional to the maximum value attained by ), and K is the *carrying capacity*. As our simulations produce the complete time series for P, we can estimate  using finite differences, thereby extracting parameters r and K through a linear regression as equation 1 suggests. In the regression, the independent and dependent variables are P and  / P, respectively. In addition, we measured the goodness of fit as that of the linear regression. Since the simulation is stochastic, multiple replications are needed for each configuration to obtain an average behavior. We used the CI method [54] to perform enough replications so that for every time step t, the 95% CI of P at time t falls within 5% of the average. Therefore, we performed the fitting for each individual run and computed the average r and K across all runs.

Although we report the carrying capacity K in Multimedia Appendices 1 and 2, the interpretation of this variable can be difficult for a broader audience. The growth rate r is *proportional* to the maximum *fraction of the carrying capacity* K that is infected on the worst day. In other words, it is an indication of how fast the disease spreads at its peak, based on another variable. For ease of interpretation, we focused on the adjusted growth rate whose unit is directly in number of individuals. The adjusted growth rate reported in this paper is obtained as:

For instance, an adjusted value of 200,000 means that at most 200,000 individuals will be infected on the worst day.

As the early steps of the simulation witness a shift from a vaccine-naïve population to one that gradually builds vaccine-based immunity, early trends differ from the longer ones that are the focus of this study. This is a typical situation in modeling, whereby estimating the long run performance measures requires to first run the model for a certain amount of time (known as the *warm-up period*) [55]. We empirically determined that a warm-up period of 20 days was sufficient to start the curve fitting; that is, we created the time series for P starting from $t \geq 20$. As evidenced by Figure 4, this warm-up period results in very good fit for the logistic model under both federal plans. This approach also generalizes better, since the reported r and K can accurately characterize the spread of the disease for most time periods instead of being skewed by the first few days.

An essential aspect of a return to normalcy is about the *conditions* under which that is achieved. If the disease is left uncontrolled, and simplifying the matter of variants, we would still return to *normalcy* within 6 months because a large share of the population would already have been infected and either recovered or died (Figure 5). The goal is thus not *only* to eventually achieve stability in the number of cases but to achieve it at a minimal level (Figure 5; bottom blue curve).

Figure 4. Distributions of the average goodness of fit R^2 for each vaccination plan, demonstrating the validity of fitting logistic growth models from t≥20.

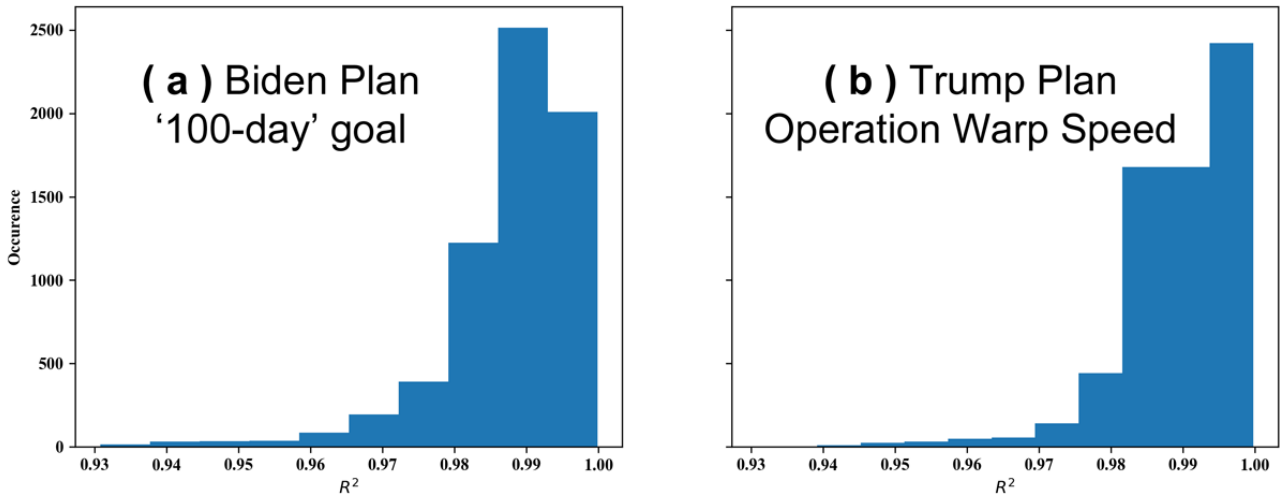
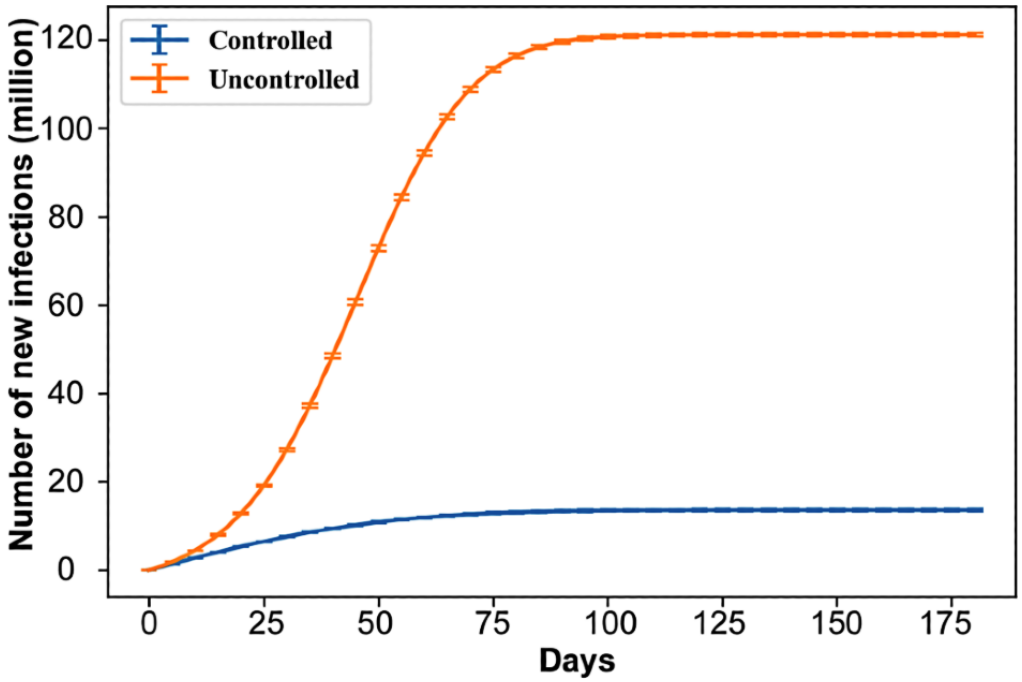


Figure 5. Number of new infections during the simulation (ie, cumulative cases) under Operation Warp Speed with vaccine compliance of 0.6, vaccine efficacy of 0.99, scenario 1 for nonpharmaceutical interventions ("controlled" case: blue), and scenario 6 consisting of no interventions ("uncontrolled" case: orange).



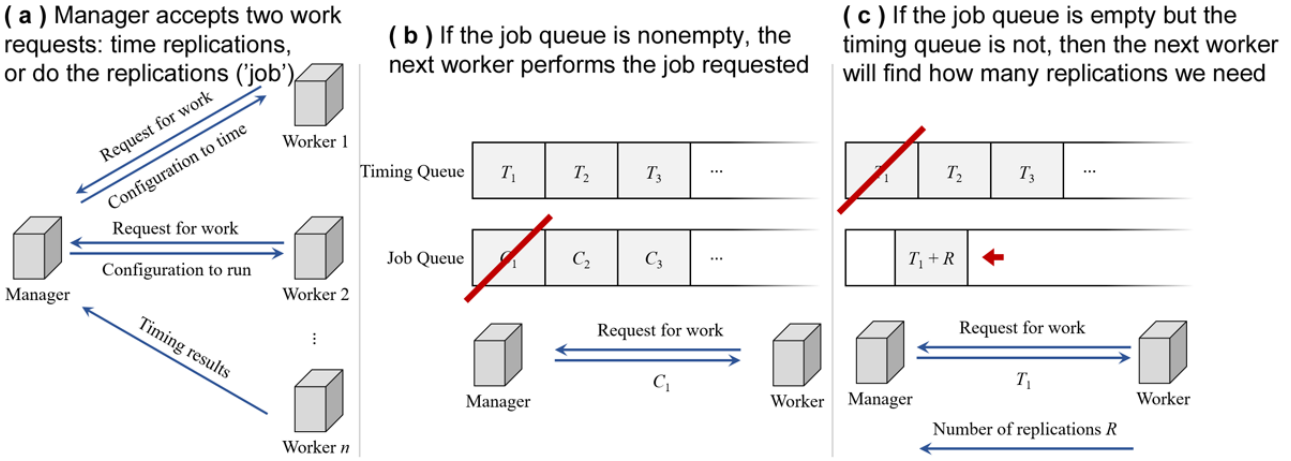
Simulation Management in Azure

To efficiently orchestrate simulations over the Microsoft Azure cloud computing platform, we used a distributed scheme shown in Figure 6. The setup starts by creating a manager, which uses queues to organize the two types of work that need to be performed.

1. Given a configuration (eg, which scenario, compliance level, and vaccine efficacy), they need to determine how many replications are necessary for a tight CI of 95%. These tasks are tracked in the timing queue.
2. Given a configuration and set number of replications, perform the computations to produce the results. These tasks are tracked in the job queue.

Available workers contact the manager, who will assign work (Figure 6a) by prioritizing the job queue and then the timing queue. For example, if a worker notifies the manager that it is available and there is a simulation run to perform in the job queue, then the manager will hand that one run to the worker (Figure 6b). If a worker is available and all queued simulations have been performed, then the manager will task the worker with identifying how many simulations are necessary for the next configuration (Figure 6c), which will refill the job queue.

Figure 6. Our simulation management architecture to leverage parallelism on Microsoft Azure.



Results

The carrying capacities and growth rates as functions of vaccine compliance and efficacies for each vaccination plan are provided in Multimedia Appendices 1-4. In this paper, we focused on the adjusted growth rate in Figures 7 and 8 for the two federal plans, 6 scenarios (including 5 nonpharmaceutical interventions), and by varying vaccine efficacy and compliance. This allowed us to examine the synergistic effects of nonpharmaceutical interventions with vaccines while comprehensively accounting for key unknowns.

In comparing the two federal plans, the Biden plan showed more potency at controlling the infection across all intervention scenarios than the plan created under the previous administration. We noted that even if a small fraction of the population seeks vaccines, and even if vaccines are less effective than announced, the vaccination campaign can reduce the total number of infections. Note that increasing the efficacies of vaccines results in lower infections for all scenarios and vaccine plans. This agrees with expectations since, in our simulations, agents are not revaccinated upon having no immune response. Therefore, holding all else equal, increasing the vaccine efficacy accelerates the growth of the immune population, thereby reaching herd immunity more quickly. In contrast, the dependence on compliance is much less intuitive and even leads to unintended consequences.

Typically, we assumed that higher vaccine compliance will lead to lower overall infections, since the proportion of the immune population is upper bounded by the compliance. However, in both vaccination plans, only scenarios 1 and 2 yielded such results. For the rest of the scenarios (3-6), the dependence on vaccine compliance is apparently reversed, with some hinting toward a nonmonotonic relationship (scenario 4 of the Biden plan and scenario 5 of the federal plan, for example). The reason behind this puzzling behavior is a combination of three factors: (1) vaccines are strictly administered in decreasing order of age; (2) older adults are going neither to work nor to school, hence they have fewer social ties than other age groups, which reduces their impact on preventing the spread of infections once immunized; and (3) relative to the growth of infections in the scenarios in which the anomaly happen, the vaccine availabilities are too low.

If we assume that an increase in vaccine compliance at the population-level is approximately uniform across age categories, then a rising vaccine compliance means that more older adults will seek vaccines. If they are also given priority for vaccines, then an increase in vaccine compliance will lead to more doses being used by older adults, hence more time to provide access to younger age groups. In short, under a vaccination strategy that focuses on older individuals, an increase in vaccine compliance will increase the delay before the more connected and younger age groups can be vaccinated. During this time, the virus can continue to spread among the younger population, particularly because the scenarios with counterintuitive results (3-6) are among the least restrictive in terms of nonpharmaceutical interventions and older adults have a lower contribution to the spread of infections due to their more limited social ties. Therefore, although the older adult population will be better protected, the longer delay for the rest of the population means that by the time they are eligible for vaccinations, the infection has already spread, leading to overall higher infections.

This argument is most vividly illustrated by our animations in Multimedia Appendices 3 and 4, in which the distributions of the infected and immune population are plotted at each time step. These animations showcase the no-intervention scenario (scenario 6) and Operation Warp Speed for the monthly vaccination capacity. Apart from the compliance, every other parameter including the random seed is fixed to be the same. Particular attention should be paid to the spread of infection among the older adult agents (ie, 65 years and older), as it most directly corroborates the aforementioned reasoning.

Figure 7. Adjusted growth rate (number of infected individuals on the worst day) as functions of vaccine compliance and efficacy under the Biden vaccination plan.

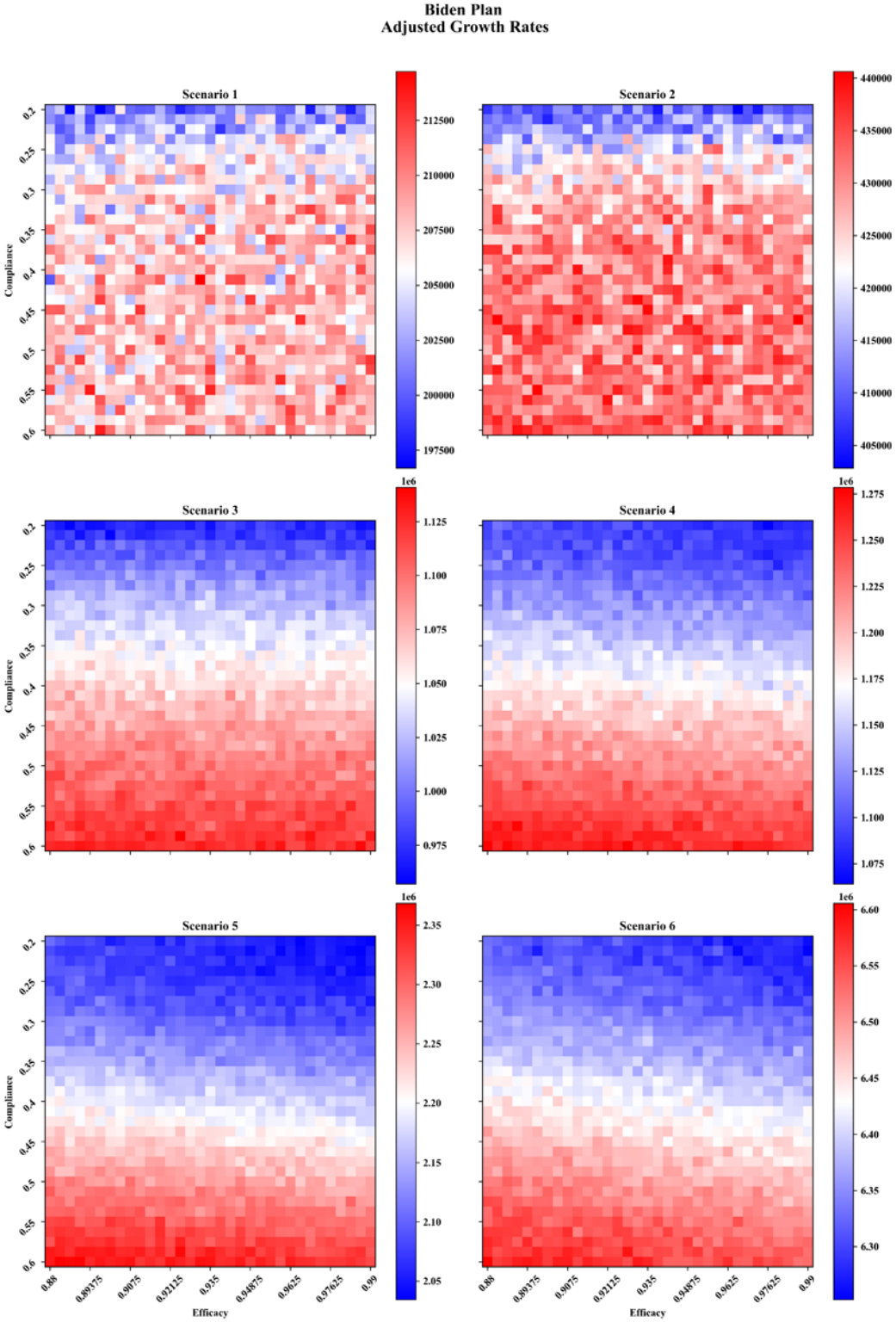
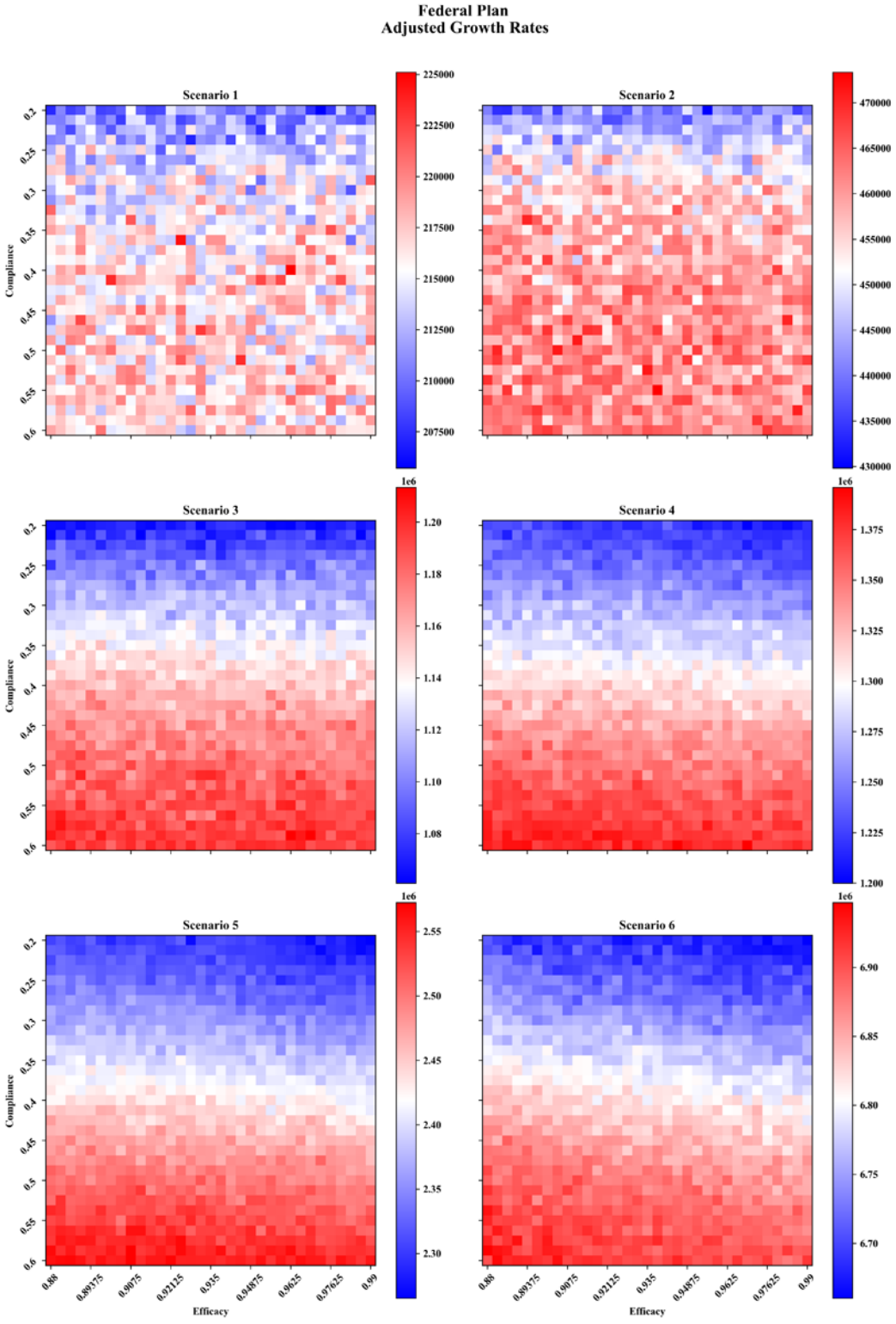


Figure 8. Adjusted growth rate (number of infected individuals on the worst day) as functions of vaccine compliance and efficacy under the Trump vaccination plan.



Discussion

Principal Results

The incoming CDC director predicted half a million deaths by mid-February 2021 [56], thus stressing the urgency of vaccination. However, vaccination is an unprecedented and complex endeavor whose success depends on many other variables such as vaccine compliance, vaccine efficacy, and the ongoing presence of nonpharmaceutical interventions. In line with expectations, our large-scale agent-based simulations showed that vaccination can reduce the total number of infections across all possible scenarios. The capacity pledged under the new Biden plan (one million doses a day) would have a greater impact than the plan of the previous administration (*Operation Warp Speed*) when accounting for its initial delays.

Two key findings of our study are as follows. First, we demonstrated the necessity to maintain nonpharmaceutical interventions over the next 6 months. As interventions are relaxed (from scenario 1 offering the most control to scenario 6 offering no control), there is an increase in case count such that a return to normalcy is not achieved through vaccination but rather through a very high number of infected individuals. Second, there is an unexpected interplay between vaccination strategies, nonpharmaceutical interventions, and vaccination availabilities. As nonpharmaceutical interventions lose momentum (scenarios 3 and above), an increase in vaccine compliance leads to an unexpected increase in infections due in part on the low availability of vaccines and the priority on vaccinating older adults. More so than the observation that tighter nonpharmaceutical interventions result in the slower spread of infections, this result further delineates the necessity of preparing the population to continuing nonpharmaceutical interventions even as the vaccination progresses.

Limitations

There are three main limitations to our current understanding of the COVID-19 pandemic and the vaccination campaign that affect how our simulations account for (1) the number of vaccines that *can* be administered each month, (2) biological aspects, and (3) *healthy* or asymptomatic carriers.

First, an unprecedented vaccine campaign comes with logistical challenges and uncertainty given the complex array of factors involved. As a result, fewer than the expected number of doses may be administered: federal officials aimed at giving the first dose to 20 million people during December 2020, but various delays resulted in fewer than 3 million people receiving a first dose [57]. It was recently reported that "federal officials say they do not fully understand the cause of the delays" [57] and that the administration "pledged to immediately distribute millions of COVID-19 vaccine doses from a stockpile that the U.S. health secretary has since acknowledged does not exist" [58]. This situation has resulted in views that "much of the narrative earlier this year regarding Warp Speed's preparation appears to be a sham" [59], reinforced by reports that the Biden administration found no vaccine distribution plan upon taking over from their predecessors [60]. Some of the factors causing a delay are known: there can be shipping delays or delays in administering doses due to a lack of hospital staff members, as they are already caring for individuals infected with COVID-19. Other factors may be more surprising, such as the intentional destruction of vaccine doses by hospital staff [61]. As any simulation model is necessarily a simplification, we did not include factors whose value would be entirely unknown (eg, what will be the shipment delay?) or whose existence is anecdotal given the total number of doses (eg, intentional destruction or storage errors). We were limited in our ability to use real-world numbers on how many individuals received the vaccine, as this data is captured at the state level, and several states' reporting systems have experienced errors [62]. Although there are efforts at centralizing data (eg, national news outlets aggregate data across states [63]), the level and nature of errors differ across states, which is a challenge to estimate overall model uncertainty.

We have thus followed the federal plan for the number of individuals who can get vaccinated each month. Out of all the doses that are *planned*, fewer may be *distributed* and an even lower number may ultimately be *administered*. Our simulations are thus likely representing an upper bound on the number of vaccines administered, leading to *more optimistic results than in reality*. The gap is particularly pronounced in December 2020 and may remain significant in January 2021, but early logistical issues and delays should be gradually addressed, such that the gap between federal expectations and actual implementation narrows over time.

Second, all biological aspects of the virus are based on the strains that dominated throughout 2020. Epidemiological studies from these strains have informed parameters such as transmissibility, incubation period, the proportion of asymptomatic carriers, the severity of symptoms and hence the course of the disease, and the efficacy of treatments or vaccines. The existence of different strains is well established, as phylogenies have shown seven distinct lineages [64,65], but there has not yet been a documented need to ascribe different parameter values (ie, different viral *behaviors*) to each strain. There are two possible reasons. First, there are relatively few mutations and thus a limited *chance* of a drastically different outcome naturally occurring: the virus is "considered a slowly-evolving virus as it possesses an inherent proofreading mechanism to repair the mismatches during its replication" [65]. Second, there has been little selective pressure on the virus, as it was spreading through a population that had never been exposed to an antigen (ie, immunologically naïve). Both arguments are now changing.

A new strain from the lineage B.1.1.7, named Variant of Concern 202012/01 (denoted VOC-202012/01), emerged with an unusually large number of 23 changes in its genomes (including mutations and deletions) [66]. Some of the biological changes make it easier for the virus to attach to its targets and enter cells, which is captured through epidemiological indicators as increased transmissibility [67,68]. This is relevant for our study, as this more contagious COVID-19 strain has been spreading in the United States and may dominate by March 2020 [69]. To date, there is no peer-reviewed evidence of an impact on disease severity or vaccine efficacy over a large population sample, but the function for some of the mutated parts remains unknown (hence the possibility of an impact on severity), and early studies over 20 volunteers suggest that antibodies from vaccines are only one-third as effective on some variants [70]. In addition, vaccination means that the virus is no longer spreading through an immunologically naïve population, thus creating selective pressure for functional mutations that can help the virus adapt. Our *simulation results are thus optimistic* as they use a lower transmissibility than provided by the new strain, and we did not worsen any of the other parameters to account for possible selective pressure.

Third, our model considers that individuals who were successfully immunized can act as a buffer in the spread of the epidemic. Reality may be more nuanced, as viral transmission from a vaccinated host to an unvaccinated one may be possible. At the time of writing (March 2021), we do not yet have conclusive findings about this possibility. As trials continue,

we may find that immunized individuals should be treated in a model as *healthy carriers* for a period. We also noted that the immunity conferred by the vaccine appears to have a different response than the immunity acquired by recovering from a natural infection. That is, a vaccine promotes the production of antibodies in the blood, but a natural immunity may lead to developing antibodies in the mucosal regions [71], which are the first site of infection (in the nose and mouth). From a modeling viewpoint, the two immunities may thus have to be treated differently in the future.

Finally, we note that our model is *built specifically for the United States*. It would not be accurate when transposed to another country with minimal changes (eg, only reducing the population size). For example, stark differences in vaccine rollout strategies exist between the United Kingdom and the United States, which would affect our simulations. In the United States, two doses of the same vaccine are normally administered, as the CDC stated that "mRNA COVID-19 vaccines are *not* interchangeable with each other or with other COVID-19 vaccine products" [72]. However, new guidance from the United Kingdom allows a mix-and-match vaccine regimen in which the second dose may be from a *different* vaccine in exceptional circumstances (eg, if the vaccine from the first dose is not available upon the patient's return), even though clinical trials for mixed regimens are only due to be conducted at a later, unspecified time [73]. Another difference is that the United Kingdom front-loads the vaccine by delivering *as many first doses* as possible, which thus no longer guarantees that a patient can receive the corresponding second dose upon return (hence raising the need for a mix-and-match) and potentially delays the delay before a second dose up to 12 weeks [73]. In contrast, the United States is against delaying the second dose [74], thus our model operates on the assumption that a patient can complete treatment on time.

Related Works: The Scale of Agent-Based Models for COVID-19

Our simulation of half a million agents qualifies as large-scale *in the context* of COVID-19 ABMs. In another context, the scale may be different as the computational costs of the simulation or historical practices in a research community may differ. For example, in HIV research, simulations have used half a million cells for about 20 years on personal computers, so a *large-scale* may be a more appropriate qualifier for simulations with billion cells [75,76]. As noted by Gumel and colleagues [77] in their extensive discussion on modeling methods for COVID-19, ABMs "are computationally-intensive"; thus, we may expect a smaller simulated population than in compartmental models or meta-population models, given the same hardware and simulation time.

Many ABMs for COVID-19 are in the scale of several hundred agents [78-83] to tens of thousands of agents [37,84,85]. Fewer studies have over 100,000 agents [86], and only a paucity of studies has a number of agents that is about equal (eg, the model of Hoertel and colleagues [87] used 500,000 agents) or greater than (eg, one million agents in a February 2021 simulation of Bogota) in this study [38,87,88]. Due to this distribution of agent population across studies, the qualifier of *large* is applied as we get to the scale of 500,000 or more agents [38]. It should not be interpreted to suggest that this is the *largest* population size achieved to date. Indeed, a few high-profile studies have modeled their target populations with such a fine resolution that the simulation may qualify as a *digital twin*. For example, Chang et al [89] used over 24 million agents by adding a COVID-19 component (AMTraC-19) to an existing model and running it over 4264 compute cores.

Although a justification for the scale is a recommended best practice in ABM for artificial societies [90], such a justification is not always present in published studies. The studies that have justified their choice of scale have often done it based on the size of the target population (eg, single city or campus) or implicitly invoked the notion of a computational burden when downscaling. Explicit mentions of computational costs have been made by the developers of frameworks, such as Comokit, who stated that 10-20,000 agents could be simulated on one laptop within 10 minutes [37].

Conclusions

A desirable return to normalcy would be achieved via immunization rather than through a very high number of infected cases and their natural immunity. Our extended ABM shows that vaccines are not sufficient to return to normalcy while avoiding a high number of cases. Nonpharmaceutical interventions are necessary and require a high level of compliance to ensure that immunity from vaccination outpaces the immunity from infections. Although our findings account for different vaccination capabilities, compliance levels, and vaccine efficacy, they are nonetheless based on a simulation model, which is necessarily a simplification of reality. Simplifications here include the logistics of vaccine dissemination, variants, and the presence of *healthy carriers* (vaccinated) and asymptomatic cases (not vaccinated).

Acknowledgments

The authors thank the Microsoft AI for Health program for supporting this research effort through a philanthropic grant. The sponsor did not influence the design, methods, or analyses in this study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Biden carrying capacity.

[PNG File , 455 KB - medinform_v9i4e27419_app1.png]

Multimedia Appendix 2
Operation Warp Speed carrying capacity.
[PNG File , 475 KB - medinform_v9i4e27419_app2.png]

Multimedia Appendix 3
High compliance.
[MP4 File (MP4 Video), 120 KB - medinform_v9i4e27419_app3.mp4]

Multimedia Appendix 4
Low compliance.
[MP4 File (MP4 Video), 121 KB - medinform_v9i4e27419_app4.mp4]

References

1. COVID Data Tracker. Centers for Disease Control and Prevention. URL: https://covid.cdc.gov/covid-data-tracker/#cases_casesper100klast7days [accessed 2021-01-21]
2. Reese H, Iuliano AD, Patel NN, Garg S, Kim L, Silk BJ, et al. Estimated incidence of COVID-19 illness and hospitalization - United States, February-September, 2020. Clin Infect Dis 2020 Nov 25:ciaa1780 [FREE Full text] [doi: 10.1093/cid/ciaa1780] [Medline: 33237993]
3. Rolfes MA, Foppa IM, Garg S, Flannery B, Brammer L, Singleton JA, et al. Annual estimates of the burden of seasonal influenza in the United States: a tool for strengthening influenza surveillance and preparedness. Influenza Other Respir Viruses 2018 Jan;12(1):132-137. [doi: 10.1111/irv.12486] [Medline: 29446233]
4. Andrasfay T, Goldman N. Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations. Proc Natl Acad Sci U S A 2021 Feb 02;118(5):e2014746118. [doi: 10.1073/pnas.2014746118] [Medline: 33446511]
5. Morlacco A, Motterle G, Zattoni F. The multifaceted long-term effects of the COVID-19 pandemic on urology. Nat Rev Urol 2020 Jul;17(7):365-367 [FREE Full text] [doi: 10.1038/s41585-020-0331-y] [Medline: 32377015]
6. Stobbe M. US deaths in 2020 top 3 million, by far most ever counted. Associated Press News. 2020 Dec 22. URL: https://apnews.com/article/us-coronavirus-deaths-top-3-million-e2bc856b6ec45563b84ee2e87ae8d5e7 [accessed 2021-01-21]
7. Fraser E. Long term respiratory complications of covid-19. BMJ 2020 Aug 03;370:m3001. [doi: 10.1136/bmj.m3001] [Medline: 32747332]
8. Rimmer A. Covid-19: Impact of long term symptoms will be profound, warns BMA. BMJ 2020 Aug 13;370:m3218. [doi: 10.1136/bmj.m3218] [Medline: 32816748]
9. Blair PW, Brown D, Jang M, Antar AAR, Keruly JC, Bachu VS, et al. The clinical course of COVID-19 in the outpatient setting: a prospective cohort study. medRxiv Preprint posted online on September 3, 2020. [doi: 10.1093/ofid/ofab007] [Medline: 33614816]
10. Becker RC. Anticipating the long-term cardiovascular effects of COVID-19. J Thromb Thrombolysis 2020 Oct;50(3):512-524 [FREE Full text] [doi: 10.1007/s11239-020-02266-6] [Medline: 32880795]
11. Mitrani RD, Dabas N, Goldberger JJ. COVID-19 cardiac injury: implications for long-term surveillance and outcomes in survivors. Heart Rhythm 2020 Nov;17(11):1984-1990 [FREE Full text] [doi: 10.1016/j.hrthm.2020.06.026] [Medline: 32599178]
12. Shaw B, Daskareh M, Gholamrezanezhad A. The lingering manifestations of COVID-19 during and after convalescence: update on long-term pulmonary consequences of coronavirus disease 2019 (COVID-19). Radiol Med 2021 Jan;126(1):40-46 [FREE Full text] [doi: 10.1007/s11547-020-01295-8] [Medline: 33006087]
13. Lu Y, Li X, Geng D, Mei N, Wu P, Huang C, et al. Cerebral micro-structural changes in COVID-19 patients - an MRI-based 3-month follow-up study. EClinicalMedicine 2020 Aug;25:100484 [FREE Full text] [doi: 10.1016/j.eclinm.2020.100484] [Medline: 32838240]
14. Pradhan D, Biswasroy P, Kumar Naik P, Ghosh G, Rath G. A review of current interventions for COVID-19 prevention. Arch Med Res 2020 Jul;51(5):363-374 [FREE Full text] [doi: 10.1016/j.arcmed.2020.04.020] [Medline: 32409144]
15. Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, et al. Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. Cochrane Database Syst Rev 2020 Apr 08;4:CD013574 [FREE Full text] [doi: 10.1002/14651858.CD013574] [Medline: 32267544]
16. Cheng C, Barceló J, Hartnett AS, Kubinec R, Messerschmidt L. COVID-19 government response event dataset (CoronaNet v.1.0). Nat Hum Behav 2020 Jul;4(7):756-768. [doi: 10.1038/s41562-020-0909-7] [Medline: 32576982]
17. Capano G, Howlett M, Jarvis DS, Ramesh M, Goyal N. Mobilizing policy (in)capacity to fight COVID-19: understanding variations in state responses. Policy Soc 2020 Jul 03;39(3):285-308. [doi: 10.1080/14494035.2020.1787628]

18. Different COVID-19 vaccines. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines.html [accessed 2021-01-21]
19. Mascarenhas L, Maxouris C, Levenson E, Almasy S. Fauci says US can return to normal by fall if it puts aside slow start and is diligent about vaccinations. CNN. URL: https://www.cnn.com/2020/12/30/health/us-coronavirus-wednesday/index.html [accessed 2021-01-21]
20. He S, Peng Y, Sun K. SEIR modeling of the COVID-19 and its dynamics. Nonlinear Dyn 2020 Jun 18:1-14 [FREE Full text] [doi: 10.1007/s11071-020-05743-y] [Medline: 32836803]
21. Yang Z, Zeng Z, Wang K, Wong S, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis 2020 Mar;12(3):165-174. [doi: 10.21037/jtd.2020.02.64] [Medline: 32274081]
22. Annas S, Isbar Pratama M, Rifandi M, Sanusi W, Side S. Stability analysis and numerical simulation of SEIR model for pandemic COVID-19 spread in Indonesia. Chaos Solitons Fractals 2020 Oct;139:110072 [FREE Full text] [doi: 10.1016/j.chaos.2020.110072] [Medline: 32834616]
23. Khyar O, Allali K. Global dynamics of a multi-strain SEIR epidemic model with general incidence rates: application to COVID-19 pandemic. Nonlinear Dyn 2020 Sep 08:1-21 [FREE Full text] [doi: 10.1007/s11071-020-05929-4] [Medline: 32921921]
24. Chang MC, Park Y, Kim B, Park D. Risk factors for disease progression in COVID-19 patients. BMC Infect Dis 2020 Jun 23;20(1):445 [FREE Full text] [doi: 10.1186/s12879-020-05144-x] [Medline: 32576139]
25. Jordan RE, Adab P, Cheng KK. Covid-19: risk factors for severe disease and death. BMJ 2020 Mar 26;368:m1198. [doi: 10.1136/bmj.m1198] [Medline: 32217618]
26. Harper CA, Satchell LP, Fido D, Latzman RD. Functional fear predicts public health compliance in the COVID-19 pandemic. Int J Ment Health Addict 2020 Apr 27:1-14 [FREE Full text] [doi: 10.1007/s11469-020-00281-5] [Medline: 32346359]
27. Nivette A, Ribeaud D, Murray A, Steinhoff A, Bechtiger L, Hepp U, et al. Non-compliance with COVID-19-related public health measures among young adults in Switzerland: insights from a longitudinal cohort study. Soc Sci Med 2021 Jan;268:113370 [FREE Full text] [doi: 10.1016/j.socscimed.2020.113370] [Medline: 32980677]
28. Snyder BF, Parks V. Spatial variation in socio-ecological vulnerability to Covid-19 in the contiguous United States. Health Place 2020 Nov;66:102471. [doi: 10.1016/j.healthplace.2020.102471] [Medline: 33129050]
29. Henning-Smith C. The unique impact of COVID-19 on older adults in rural areas. J Aging Soc Policy 2020;32(4-5):396-402. [doi: 10.1080/08959420.2020.1770036] [Medline: 32475255]
30. Amram O, Amiri S, Lutz RB, Rajan B, Monsivais P. Development of a vulnerability index for diagnosis with the novel coronavirus, COVID-19, in Washington State, USA. Health Place 2020 Jul;64:102377 [FREE Full text] [doi: 10.1016/j.healthplace.2020.102377] [Medline: 32838894]
31. Karim SA, Chen H. Deaths from COVID-19 in rural, micropolitan, and metropolitan areas: a county-level comparison. J Rural Health 2021 Jan;37(1):124-132. [doi: 10.1111/jrh.12533] [Medline: 33155723]
32. Pollmann TR, Pollmann J, Wiesinger C, Haack C, Shtembari L, Turcati A, et al. The impact of digital contact tracing on the SARS-CoV-2 pandemic - a comprehensive modelling study. medRxiv Preprint posted online on September 14, 2020. [doi: 10.1101/2020.09.13.20192682]
33. Firth JA, Hellewell J, Klepac P, Kissler S, CMMID COVID-19 Working Group, Kucharski AJ, et al. Using a real-world network to model localized COVID-19 control strategies. Nat Med 2020 Oct;26(10):1616-1622. [doi: 10.1038/s41591-020-1036-8] [Medline: 32770169]
34. Li J, Giabbanelli PJ. Identifying synergistic interventions to address COVID-19 using a large scale agent-based model. medRxiv Preprint posted online on December 14, 2020. [doi: 10.1101/2020.12.11.20247825]
35. Kerr CC, Stuart RM, Mistry D, Abeysuriya RG, Rosenfeld K, Hart GR, et al. Covasim: an agent-based model of COVID-19 dynamics and interventions. medRxiv Preprint posted online on April 1, 2021. [doi: 10.1101/2020.05.10.20097469]
36. Hinch R, Probert WJM, Nurtay A, Kendall M, Wymant C, Hall M, et al. OpenABM-Covid19 - an agent-based model for non-pharmaceutical interventions against COVID-19 including contact tracing. medRxiv Preprint posted online on September 22, 2020. [doi: 10.1101/2020.09.16.20195925]
37. Gaudou B, Huynh NQ, Philippon D, Brugière A, Chapuis K, Taillandier P, et al. COMOKIT: a modeling kit to understand, analyze, and compare the impacts of mitigation policies against the COVID-19 epidemic at the scale of a city. Front Public Health 2020;8:563247. [doi: 10.3389/fpubh.2020.563247] [Medline: 33072700]
38. Aleman DM, Tham B, Wagner SJ, Semelhago J, Mohammadi A, Price P, et al. Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering. 2020 Presented at: CASCON '20; November 10-13, 2020; Online p. 266-267. [doi: 10.1101/2021.02.05.21251157]
39. Panovska-Griffiths J, Kerr CC, Stuart RM, Mistry D, Klein DJ, Viner RM, et al. Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a second COVID-19 epidemic wave in the UK: a modelling study. Lancet Child Adolesc Health 2020 Nov;4(11):817-827 [FREE Full text] [doi: 10.1016/S2352-4642(20)30250-9] [Medline: 32758453]

40. Scott N, Palmer A, Delport D, Abeysuriya R, Stuart RM, Kerr CC, et al. Modelling the impact of relaxing COVID-19 control measures during a period of low viral transmission. Med J Aust 2021 Feb;214(2):79-83 [FREE Full text] [doi: 10.5694/mja2.50845] [Medline: 33207390]
41. Yehya N, Venkataramani A, Harhay MO. Statewide interventions and covid-19 mortality in the United States: an observational study. Clin Infect Dis 2020 Jul 08:ciaa923 [FREE Full text] [doi: 10.1093/cid/ciaa923] [Medline: 32634828]
42. Lewnard JA, Lo NC. Scientific and ethical basis for social-distancing interventions against COVID-19. Lancet Infect Dis 2020 Jun;20(6):631-633 [FREE Full text] [doi: 10.1016/S1473-3099(20)30190-0] [Medline: 32213329]
43. Lazer D, Santillana M, Perlis RH, Ognyanova K, Baum MA. State of the Nation: a 50-state COVID-19 survey: report #8: failing the test: waiting times for COVID diagnostic tests across the U.S. Pesquisa. URL: https://search.bvsalud.org/global-literature-on-novel-coronavirus-2019-ncov/resource/en/grc-740424 [accessed 2021-01-21]
44. Klein B, Stracqualursi V, Sullivan K. Biden unveils Covid-19 plan based on 'science not politics' as he signs new initiatives. CNN. URL: https://www.cnn.com/2021/01/21/politics/biden-national-coronavirus-plan/index.html [accessed 2021-01-21]
45. Vera A, Watts A, Langmaid V, Holcombe M. US reports over 200K new Covid-19 cases every single day for a week straight. CNN. URL: https://www.cnn.com/2021/01/11/health/us-coronavirus-monday/index.html [accessed 2021-01-21]
46. Israel's virus czar says 1st dose less effective than Pfizer indicated — report. The Times of Israel. 2021 Jan 19. URL: https://www.timesofisrael.com/israels-virus-czar-says-1st-dose-less-effective-than-pfizer-indicated-report/ [accessed 2021-01-21]
47. Wang X. Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. N Engl J Med 2021 Apr 22;384(16):1577-1578. [doi: 10.1056/NEJMc2036242] [Medline: 33596350]
48. Khubchandani J, Sharma S, Price JH, Wiblishauser MJ, Sharma M, Webb FJ. COVID-19 vaccination hesitancy in the United States: a rapid national assessment. J Community Health 2021 Apr;46(2):270-277 [FREE Full text] [doi: 10.1007/s10900-020-00958-x] [Medline: 33389421]
49. Lin C, Tu P, Beitsch LM. Confidence and receptivity for COVID-19 vaccines: a rapid systematic review. Vaccines (Basel) 2020 Dec 30;9(1):16 [FREE Full text] [doi: 10.3390/vaccines9010016] [Medline: 33396832]
50. Pogue K, Jensen JL, Stancil CK, Ferguson DG, Hughes SJ, Mello EJ, et al. Influences on attitudes regarding potential COVID-19 vaccination in the United States. Vaccines (Basel) 2020 Oct 03;8(4):582 [FREE Full text] [doi: 10.3390/vaccines8040582] [Medline: 33022917]
51. Malavika B, Marimuthu S, Joy M, Nadaraj A, Asirvatham ES, Jeyaseelan L. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic growth models. Clin Epidemiol Glob Health 2021;9:26-33 [FREE Full text] [doi: 10.1016/j.cegh.2020.06.006] [Medline: 32838058]
52. Kamrujjaman M, Mahmud MS, Islam MS. Coronavirus outbreak and the mathematical growth map of COVID-19. Annu Res Rev Biol 2020 Mar 26:72-78. [doi: 10.9734/arrb/2020/v35i130182]
53. Wang P, Zheng X, Li J, Zhu B. Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics. Chaos Solitons Fractals 2020 Oct;139:110058 [FREE Full text] [doi: 10.1016/j.chaos.2020.110058] [Medline: 32834611]
54. Robinson S. Simulation: The Practice of Model Development and Use. London, UK: Red Globe Press; 2001.
55. Mahajan PS, Ingalls RG. Evaluation of methods used to detect warm-up period in steady state simulation. 2004 Presented at: 2004 Winter Simulation Conference; December 5-8, 2004; Washington, DC. [doi: 10.1109/wsc.2004.1371374]
56. Caldwell T, Yan H. Incoming CDC director: expect 500,000 Covid-19 deaths by mid-February. CNN. URL: https://www.cnn.com/2021/01/17/health/us-coronavirus-sunday/index.html [accessed 2021-01-21]
57. Robbins R, Robles F, Arangon T. Here's why distribution of the vaccine is taking longer than expected. The New York Times. 2020 Dec 31. URL: https://www.nytimes.com/2020/12/31/health/vaccine-distribution-delays.html [accessed 2021-01-21]
58. Szekely P, Gorman S. Trump administration accused of deception in pledging release of vaccine stockpile. Reuters. URL: https://www.reuters.com/article/us-health-coronavirus-usa/trump-administration-accused-of-deception-in-pledging-release-of-vaccine-stockpile-idUSKBN29K2H4 [accessed 2021-01-21]
59. Kavanagh K. With the coronavirus mutating and vaccinations behind schedule, here's what we must do now. Courier Journal. 2021 Jan 08. URL: https://www.courier-journal.com/story/opinion/2021/01/08/with-coronavirus-vaccinations-behind-schedule-heres-what-us-must-do/6557584002/ [accessed 2021-01-21]
60. Lee MJ. Biden inheriting nonexistent coronavirus vaccine distribution plan and must start 'from scratch,' sources say. CNN. URL: https://www.cnn.com/2021/01/21/politics/biden-covid-vaccination-trump/index.html [accessed 2021-01-21]
61. Romo V. Some 500 Coronavirus vaccine doses intentionally destroyed, hospital says. NPR. 2020 Dec 31. URL: https://www.npr.org/2020/12/30/951736164/some-500-coronavirus-vaccine-doses-intentionally-destroyed-hospital-says [accessed 2021-01-21]
62. Harris C, Picon A, Despart Z. As Texas leaders claim COVID vaccines are sitting on shelves, hospitals and pharmacies beg for more. Houston Chronicle. 2020 Dec 31. URL: https://www.houstonchronicle.com/news/houston-texas/health/article/texas-leaders-covid-vaccines-hospitals-pharmacy-15838452.php [accessed 2021-01-21]
63. Covid vaccine, states distribution soses. The Washington Post. URL: https://www.washingtonpost.com/graphics/2020/health/covid-vaccine-states-distribution-doses/ [accessed 2021-01-21]

64. Kumar S, Tao Q, Weaver S, Sanderford M, Caraballo-Ortiz MA, Sharma S, et al. An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. bioRxiv Preprint posted online on January 19, 2021. [doi: 10.1101/2020.09.24.311845] [Medline: 32995781]
65. Jacob JJ, Vasudevan K, Pragasam AK, Gunasekaran K, Kang G, Veeraraghavan B, et al. Evolutionary tracking of SARS-CoV-2 genetic variants highlights intricate balance of stabilizing and destabilizing mutations. bioRxiv Preprint posted online on December 22, 2020.
66. Kemp SA, Meng B, Ferriera IATM, Datir R, Harvey WT, Papa G, The COVID-19 Genomics UK (COG-UK) Consortium, et al. Recurrent emergence and transmission of a SARS-CoV-2 spike deletion H69/V70. bioRxiv Preprint posted online on March 8, 2021.
67. Science brief: emerging SARS-CoV-2 variants. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/coronavirus/2019-ncov/more/science-and-research/scientific-brief-emerging-variants.html [accessed 2021-01-21]
68. Davies NG, Abbott S, Barnard RC, Jarvis CI, Kucharski AJ, Munday J, CMMID COVID-19 Working Group, et al. Estimated transmissibility and severity of novel SARS-CoV-2 Variant of Concern 202012/01 in England. medRxiv Preprint posted online on February 7, 2021. [doi: 10.1101/2020.12.24.20248822]
69. Galloway SE, Paul P, MacCannell DR, Johansson MA, Brooks JT, MacNeil A, et al. Emergence of SARS-CoV-2 B.1.1.7 lineage - United States, December 29, 2020-January 12, 2021. MMWR Morb Mortal Wkly Rep 2021 Jan 22;70(3):95-99. [doi: 10.15585/mmwr.mm7003e2] [Medline: 33476315]
70. Wang Z, Schmidt F, Weisblum Y, Muecksch F, Barnes CO, Finkin S, et al. mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants. bioRxiv Preprint posted online on January 30, 2021. [doi: 10.1101/2021.01.15.426911] [Medline: 33501451]
71. Cervia C, Nilsson J, Zurbuchen Y, Valaperti A, Schreiner J, Wolfensberger A, et al. Systemic and mucosal antibody responses specific to SARS-CoV-2 during mild versus severe COVID-19. J Allergy Clin Immunol 2021 Feb;147(2):545-557.e9 [FREE Full text] [doi: 10.1016/j.jaci.2020.10.040] [Medline: 33221383]
72. Interim clinical considerations for use of COVID-19 vaccines currently authorized in the United States. Centers for Disease Control and Prevention. URL: https://www.cdc.gov/vaccines/covid-19/info-by-product/clinical-considerations.html [accessed 2021-01-21]
73. Wu KJ. Britain opens door to mix-and-match vaccinations, worrying experts. The New York Times. 2021 Jan 01. URL: https://www.nytimes.com/2021/01/01/health/coronavirus-vaccines-britain.html [accessed 2021-01-21]
74. Mohamed E. Dr Anthony Fauci says US will not delay second doses of Covid vaccine. The Guardian. 2021 Jan 01. URL: https://www.theguardian.com/world/2021/jan/02/dr-anthony-fauci-says-us-will-not-delay-second-doses-of-covid-vaccine [accessed 2021-01-21]
75. Giabbanelli PJ, Devita JA, Köster T, Kohrt JA. Optimizing discrete simulations of the spread of hiv-1 to handle billions of cells on a workstation. In: Proceedings of the 2020 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation. 2020 Presented at: SIGSIM-PADS '20; June 2020; Miami, FL p. 67-78.
76. Köster T, Giabbanelli PJ, Uhrmacher A. Performance and soundness of simulation: a case study based on a cellular automaton for in-body spread of HIV. 2020 Presented at: 2020 Winter Simulation Conference (WSC); December 14-18, 2020; Orlando, FL p. 13-16.
77. Gumel A, Iboi E, Ngonghala C, Elbasha E. A primer on using mathematics to understand COVID-19 dynamics: modeling, analysis and simulations. Infect Dis Model 2021;6:148-168 [FREE Full text] [doi: 10.1016/j.idm.2020.11.005] [Medline: 33474518]
78. Wallentin G, Kaziyeva D, Reibersdorfer-Adelsberger E. COVID-19 intervention scenarios for a long-term disease management. Int J Health Policy Manag 2020 Dec 01;9(12):508-516 [FREE Full text] [doi: 10.34172/ijhpm.2020.130] [Medline: 32729281]
79. Bouchnita A, Jebrane A. A hybrid multi-scale model of COVID-19 transmission dynamics to assess the potential of non-pharmaceutical interventions. Chaos Solitons Fractals 2020 Sep;138:109941 [FREE Full text] [doi: 10.1016/j.chaos.2020.109941] [Medline: 32834575]
80. Silva P, Batista P, Lima H, Alves M, Guimarães FG, Silva R. COVID-ABS: an agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. Chaos Solitons Fractals 2020 Oct;139:110088 [FREE Full text] [doi: 10.1016/j.chaos.2020.110088] [Medline: 32834624]
81. Badham J, Barbrook-Johnson P, Caiado C, Castellani B. Justified stories with agent-based modelling for local COVID-19 planning. J Artif Soc Soc Simulation 2021;24(1):8. [doi: 10.18564/jasss.4532]
82. Alzu'bi AA, Alasal SIA, Watzlaf VJM. A simulation study of coronavirus as an epidemic disease using agent-based modeling. Perspect Health Inf Manag 2021;18(Winter):1g [FREE Full text] [Medline: 33633517]
83. Araya F. Modeling the spread of COVID-19 on construction workers: an agent-based approach. Saf Sci 2021 Jan;133:105022 [FREE Full text] [doi: 10.1016/j.ssci.2020.105022] [Medline: 33012995]
84. Shamil MS, Farheen F, Ibtehaz N, Khan IM, Rahman MS. An agent-based modeling of COVID-19: validation, analysis, and recommendations. Cognit Comput 2021 Feb 19:1-12 [FREE Full text] [doi: 10.1007/s12559-020-09801-w] [Medline: 33643473]

85. Goyal R, Hotchkiss J, Schooley RT, De Gruttola V, Martin NK. Evaluation of SARS-CoV-2 transmission mitigation strategies on a university campus using an agent-based network model. Clin Infect Dis 2021 Jan 19:ciab037 [FREE Full text] [doi: 10.1093/cid/ciab037] [Medline: 33462589]
86. Pesavento J, Chen A, Yu R, Kim JS, Kavak H, Anderson T, et al. Data-driven mobility models for COVID-19 simulation. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities. 2020 Presented at: ARIC '20; November 2020; Seattle, WA p. 29-38.
87. Hoertel N, Blachier M, Blanco C, Olfson M, Massetti M, Rico MS, et al. A stochastic agent-based model of the SARS-CoV-2 epidemic in France. Nat Med 2020 Sep;26(9):1417-1421. [doi: 10.1038/s41591-020-1001-6] [Medline: 32665655]
88. Gomez J, Prieto J, Leon E, Rodríguez A. INFEKTA-An agent-based model for transmission of infectious diseases: the COVID-19 case in Bogotá, Colombia. PLoS One 2021;16(2):e0245787 [FREE Full text] [doi: 10.1371/journal.pone.0245787] [Medline: 33606714]
89. Chang SL, Harding N, Zachreson C, Cliff OM, Prokopenko M. Modelling transmission and control of the COVID-19 pandemic in Australia. Nat Commun 2020 Nov 11;11(1):5710. [doi: 10.1038/s41467-020-19393-6] [Medline: 33177507]
90. Giabbanelli PJ, Voinov AA, Castellani B, Törnberg P. Ideal, best, and emerging practices in creating artificial societies. 2019 Presented at: 2019 Spring Simulation Conference (SpringSim); April 29-May 2, 2019; Tucson, AZ p. 1-12.

Abbreviations

ABM: agent-based model
CDC: Centers for Disease Control and Prevention

Edited by C Lovis; submitted 24.01.21; peer-reviewed by H Kavak, A Staffini; comments to author 27.02.21; revised version received 21.03.21; accepted 14.04.21; published 29.04.21.

Please cite as:
Li J, Giabbanelli P
Returning to a Normal Life via COVID-19 Vaccines in the United States: A Large-scale Agent-Based Simulation Study
JMIR Med Inform 2021;9(4):e27419
URL: https://medinform.jmir.org/2021/4/e27419
doi:10.2196/27419
PMID:33872188

Original Paper

A User-Centered Chatbot (Wakamola) to Collect Linked Data in Population Networks to Support Studies of Overweight and Obesity Causes: Design and Pilot Study

Sabina Asensio-Cuesta[1], PhD; Vicent Blanes-Selva[1], MSc; J Alberto Conejero[2], PhD; Ana Frigola[3], PhD; Manuel G Portolés[2], BSc; Juan Francisco Merino-Torres[4], PhD; Matilde Rubio Almanza[4], MD; Shabbir Syed-Abdul[5], PhD; Yu-Chuan (Jack) Li[5], PhD; Ruth Vilar-Mateo[6], PhD; Luis Fernandez-Luque[7], PhD; Juan M García-Gómez[1], PhD

[1]Instituto de Tecnologías de la Información y Comunicaciones, Universitat Politècnica de València, Valencia, Spain

[2]Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Valencia, Spain

[3]Department of Nutrition and Food Science, Universitat de València, Valencia, Spain

[4]Department of Endocrinology and Nutrition, Hospital La Fe, Universitat de València, Valencia, Spain

[5]International Center for Health Information Technology, Taipei Medical University, Taipei, Taiwan

[6]Unidad Mixta de Tic aplicadas a la reingeniería de procesos socio-sanitarios, Instituto de Investigación Sanitaria La Fe, Valencia, Spain

[7]Adhera Health Inc, Palo Alto, CA, United States

Corresponding Author:
Sabina Asensio-Cuesta, PhD
Instituto de Tecnologías de la Información y Comunicaciones
Universitat Politècnica de València
Camino de Vera s/n
Valencia, 46022
Spain
Phone: 34 96 387 70 07 ext 71846
Email: sasensio@dpi.upv.es

Abstract

Background: Obesity and overweight are a serious health problem worldwide with multiple and connected causes. Simultaneously, chatbots are becoming increasingly popular as a way to interact with users in mobile health apps.

Objective: This study reports the user-centered design and feasibility study of a chatbot to collect linked data to support the study of individual and social overweight and obesity causes in populations.

Methods: We first studied the users' needs and gathered users' graphical preferences through an open survey on 52 wireframes designed by 150 design students; it also included questions about sociodemographics, diet and activity habits, the need for overweight and obesity apps, and desired functionality. We also interviewed an expert panel. We then designed and developed a chatbot. Finally, we conducted a pilot study to test feasibility.

Results: We collected 452 answers to the survey and interviewed 4 specialists. Based on this research, we developed a Telegram chatbot named Wakamola structured in six sections: personal, diet, physical activity, social network, user's status score, and project information. We defined a user's status score as a normalized sum (0-100) of scores about diet (frequency of eating 50 foods), physical activity, BMI, and social network. We performed a pilot to evaluate the chatbot implementation among 85 healthy volunteers. Of 74 participants who completed all sections, we found 8 underweight people (11%), 5 overweight people (7%), and no obesity cases. The mean BMI was 21.4 kg/m^2 (normal weight). The most consumed foods were olive oil, milk and derivatives, cereals, vegetables, and fruits. People walked 10 minutes on 5.8 days per week, slept 7.02 hours per day, and were sitting 30.57 hours per week. Moreover, we were able to create a social network with 74 users, 178 relations, and 12 communities.

Conclusions: The Telegram chatbot Wakamola is a feasible tool to collect data from a population about sociodemographics, diet patterns, physical activity, BMI, and specific diseases. Besides, the chatbot allows the connection of users in a social network to study overweight and obesity causes from both individual and social perspectives.

(JMIR Med Inform 2021;9(4):e17503) doi:10.2196/17503

KEYWORDS

mHealth; obesity; overweight; chatbot; assessment; public health; Telegram; user-centered design; Social Network Analysis

Introduction

The percentage of overweight people has not stopped increasing worldwide since the 1980s [1]. In the United States, more than two-thirds of adults and nearly one-third of children and youth are overweight or obese [2]. According to the World Health Organization, in Europe, more than 50% of the population is overweight, and 20% is obese [3].

Obesity is a complex problem with individual, socioeconomic, and environmental factors [4]. From a social perspective, Fowler and Christakis conducted a study about the spread of obesity in a large social network over 32 years (Framingham Heart Study) [5] and found evidence of the "contagion" of obesity among people in close social circles. Indeed, the relevant finding in the study suggests that ties between friends have an even more significant effect on a person's risk of obesity than genes. A person's chances of becoming obese increased by 57% if he or she had a friend who became obese in a given interval. Moreover, for a wide variety of conditions and networks, Bahr et al [6] showed that individuals with similar BMIs would cluster together into groups.

Furthermore, chatbots, also referred to as conversational user interfaces, are gradually being adopted in mobile health (mHealth) apps [7] and serve to assess the long-term user experience [8]. A chatbot is a conversation platform that interacts with users via a chatting interface. Since its use can be facilitated by linkages with the major social network service messengers (eg, WhatsApp, Telegram), general users can easily access and receive various health services [9]. Laranjo et al [7] provide an overview of research related to conversational user interfaces in health care.

Previous studies suggest that chatbots may have the potential to contribute to obesity and overweight prevention and management [10]. In 2017, Kowatsch et al [11] designed a text-based health care chatbot to effectively support patients and health professionals in therapeutic settings beyond on-site consultations and applied to childhood obesity control. In 2018, Huang et al [12] developed a chatbot integrated in the SWITCHes app, where the chatbot helps monitor users' health; users also talk to the chatbot and get information in a real-time manner or take a bot's advice, including diet and exercise plans, in the context of healthy recommendation. In 2018, Holmes et al [10] described the design and development of a chatbot (WeightMentor), a self-help motivational tool for weight loss maintenance. In 2019, Stephens et al [13] implemented a behavioral coaching chatbot (Tess) to help support teens in a weight management program. However, chatbots can be useful not only in obesity control, monitoring, and promotion of healthy habits, as mentioned in these studies, but also as effective tools to collect data in large populations to study obesity causes and lead prevention. So far, face-to-face or online questionnaires are widely used to collect data directly from people about their weight, diet, and physical activity habits [14-17]. However, recent studies indicate that chatbots may be more attractive to users than classic questionnaires because people associate them with entertainment, social, and relational factors [8,18]. In addition, users have curiosity about what they view as a novel phenomenon [19].

Moreover, chatbots also enable the development of gamification strategies that can have a positive impact on health and wellbeing [20,21], already widely applied to online surveys [22]. Hamari defines gamification as "a process of enhancing services with (motivational) affordances in order to invoke gameful experiences and further behavioral outcomes" [23]. Concerning game mechanics, feedback and socialization aspects are recurrently employed to gamify eHealth. Social features, rewards, and progress tracking are powerful mechanics for producing positive effects on users [24]. Focusing on gamification strategies applied in chatbots, chatbots are able to implement mobile app stickiness strategies to improve user engagement, such as gaming, dexterity, responsiveness and feedback after coming in contact with the app, ease of figuring out how to operate the app, forums, multimedia display, and emotional engagement [25,26]. Siutila [27] identified tracking options in popular chatbots [28-30]: a system of points, leaderboards, achievements/badges, levels, story/theme, clear goals, feedback, rewards, progress, and challenge.

This study reports the user-centered design and feasibility study of a chatbot to collect linked data about diet, physical activity, weight, obesity risk, living area, and social network, to support research regarding individuals and social causes of obesity and overweight. Here, we describe the user-centered approach applied in the design and development of the chatbot. We also present a pilot study to test the chatbot's feasibility.

Methods

Ethics

Ethical approval was obtained for this study from the Ethical Committee of the Universitat Politècnica de València (UPV; Ethical Code: P7_12_11_2018).

Users' Needs Investigation

Applying a user-centered approach, we started the design of the chatbot by collecting potential users' expectations and preferences. We briefly expose the three parts into which we split the information collection: (1) a survey about interests and expectations, (2) an analysis of graphical preferences, and (3) a specialist panel's advice on the medical content. Further details of each of these parts can be found in Multimedia Appendix 1 [5,6,18,20,21,23-69].

First, a survey was designed including questions about sociodemographic data, self-perception of overweight, diet and physical activity, favorite colors for the app's purpose, the potential utility of the app, future use of the app, type of preferred diffusion, and desired functionalities. The survey included 13 questions in total.

Second, to investigate user preferences regarding the graphical features of the interface, wireframes were designed by design students and included in the survey to be scored on a 1 to 5 scale. Wireframes were designed following these general specifications: the appearance of the app breaks with the stigma of obesity and overweight and motivates its use; the app promotes a healthy lifestyle; the elements to be designed for each alternative were the chatbot's name, launch icon, splash, and main menu screen with preliminary options such as user's personal data, calculating risk, and suggesting healthy activities. To design the wireframes, students reviewed mHealth apps in the obesity and overweight field. No limitations were specified for the graphical or aesthetic features. To define the chatbot's colors, we completed the survey questions with research about current evidence regarding colors and their effects on people's feelings [31,32].

Third, we also formed an expert panel composed of 1 nutritionist and 3 clinicians, all of whom were endocrine specialists. After a project introduction, we addressed the panel with three research questions: (1) what data would be relevant for study of obesity and overweight, according to current knowledge, (2) if there are validated questionnaires to get these data, and (3) how obesity and overweight risk of a user could be assessed.

Chatbot's Design and Development

Based on the users' survey and expert criteria involved in the study, we decided to include six sections in the chatbot: Personal, Diet, physical activity habits (Activity), social network (Wakanet), status (Wakastatus), and project information (About Wakamola) (Figure 1; Multimedia Appendix 2). "Personal," "Diet," and "Activity" were interactive surveys based on standardized questionnaires, whereas "social network" implemented a sharing mechanism based on a sticky strategy. More importantly, we defined a novel user status assessment named Wakastatus. It was a gamified version of an obesity risk assessment based on diet, physical activity, and neighborhood (relationships) status to give feedback and motivation to the user. Moreover, the word "risk" was changed to "status" to provide a positive message for the user's obesity and overweight assessment. Further information about gamification elements in Wakamola are available in Table 1 and Multimedia Appendix 1.

Figure 1. Screenshot of Wakamola's main menu and diet section.

Table 1. Gamification strategies implemented in Wakamola.

Gamification strategy	Implementation in Wakamola
System of points (scores); goals	• Wakamola scores on a scale of 0 to 100 (the higher the better, goal 100) • Global status score (Wakastatus) • Diet score (Wakalimentation) • Activity score (Activity) • BMI score (WakaBMI) • Social network score (Wakasocial)
Socialization	Wakamola's social network
Feedback to the user	• Self-assessment of overweight and obesity risk: Wakamola's scores, BMI category, and level of obesity risk • User's network graphical representation, BMI/Wakastatus shown inside nodes, colors based on BMI/Wakastatus category
Emotional engagement	• Personification of the chatbot through the Wakamola character • Introduction of humanlike cues in Wakamola chatbot to increase users' emotional connection [18] (eyes, mouth, expressions of effort and happiness) • Use of emoji added to the Wakamola's text messages to create a more realistic and friendly conversation [35]

The Personal section includes 16 questions about weight, height, gender, age, level of education, marital status, how many people are at home, main activity (ie, study or work), zip code, sleep hours, and cigarette consumption. In addition, the chatbot asks if the user has ever received a diagnosis or is taking medication for hypertension, diabetes, high cholesterol, or cardiovascular disease. The clinicians defined these questions for further analysis regarding overweight and obesity factors.

Questions in the Diet section were adapted from the "Short questionnaire on frequency of dietary intake" [33]. In total, 51 questions regarding food types (items) and consumption frequencies are included. Diet question responses (items) were scored based on the "Spanish diet quality according to the healthy eating index," with items' scores from 1 to 10 (the higher the score of the item, the less healthy its consumption) [70] (Tables S1 and S2 in Multimedia Appendix 1).

In the Activity section with 7 questions, the short form of the International Physical Activity Questionnaire (IPAQ) has been applied to define the chatbot's questions and scoring. This IPAQ version is recommended, especially when the object of investigation is population monitoring [71].

The Wakanet section has been developed to share the Wakamola chatbot between contacts, following a sticky strategy. This is how the users' social networks and subnetworks are created to further analysis about how their social relations and habits could influence or be influenced from an overweight and obesity perspective. This section first shows a message with the user's total contacts, broken down by house, family, friends, and work contacts. Four different invitations are then created as chatbot messages to be shared with the target group of contacts: (1) people the user lives with (home), (2) friends, (3) family, and (4) work contacts. This section implements the community gamification strategy in the chatbot.

The Wakastatus section shows a normalized score calculated from previous data collected in the personal, diet, activity, and social network sections and normalized between 0 and 100; the higher the better. The Diet score is the sum of scores for each food and its consumption frequency; this score is also normalized between 0 and 100. The Activity score is calculated according to the short form of the IPAQ [71]. This result is normalized between 0 to 100 to present the final Activity score.

Additionally, we calculate the BMI score (WakaBMI) from the Personal section (weight and height), obtaining 100 points for normal weight (18.5-24.9 kg/m^2), 75 points for overweight (25-29.9 kg/m^2) or underweight (<18.5 kg/m^2), 50 points for obesity class 1 (30-34.9 kg/m^2), 25 points for obesity class 2 (35-39.9 kg/m^2), or 0 points for extreme obesity class 3 (≥40 kg/m^2) [72].

Finally, the social network score (Wakanet) is calculated based on the user number of contacts and the mean Wakastatus values of them.

Moreover, Wakamola is a multilanguage chatbot, including Spanish, English, and Catalan, allowing other languages to be easily included to the chatbot by adding corresponding dialogue file translation. The Wakamola chatbot is available in open access [73] under a Creative Commons license.

Focusing on the technical implementation, the chatbot engine of Wakamola is implemented as a Telegram bot using Python 3 [34]. Further technical details can be found in Multimedia Appendix 1.

Usability Evaluation

As part of the chatbot's user-centered development, a usability evaluation was carried out. The usability test focused on the process and the information user's understanding. The usability test was designed as a face-to-face, assisted session. As a requirement to perform the test, it was mandatory to have a smartphone with Telegram installed on it. First, to characterize the sample, participants answered questions about gender, age, Telegram experience, messaging system used, and previous knowledge and experience regarding bots. Participants were then asked to perform a set of 6 specific tasks with the chatbot.

Finally, the participants in the study responded to the System Usability Scale (SUS) questionnaire [36]. Further details about the usability evaluation can be found in Multimedia Appendix 1.

Pilot Study

To test the feasibility of the chatbot, we conducted a pilot study with 85 university students (volunteers) recruited face to face. Participants were asked to complete all of the chatbot's questions from the Personal, Diet, and Activity sections and to share the chatbot between them to build the social network. From the collected data, we obtained basic statistics from sociodemographic data, Wakamola scores, and BMI. Finally, we developed a free-access online tool [74] to perform the social network analysis by visualizing the network. The network visualization highlighted the users' BMI and Wakastatus and showed communities obtained based on Louvain algorithm [75].

Results

In this section, we show results from the users' needs research survey and expert panel and from the usability test. We then show outcomes from the pilot study.

Users' Needs Investigation Results

Participants in the survey were recruited by email invitation from the Vice-Rector for Social Responsibility to the UPV's university community (students, academy, and staff). The invitation included a brief description of the study and a link to the questionnaire. All participants who completed the questionnaire were included in the study. In total, 452 adults (197 males, 43.6%, and 255 females, 56.4%) participated in the survey for 11 days (Tables S3 and S4 in Multimedia Appendix 1). The sample was representative of the male and female composition of the university community and of a wide age range (from 18 to older than 65 years); likewise, it includes people from different lifestyles.

A high number of participants thought they were overweight (176/452, 38.9%). The perception of overweight increased with age. Most of them indicated having healthy dietary habits, including more women than men, at all ages. However, only half of the participants had regular physical activity. Moreover, almost half of them (217/452, 48.0%) thought that with their current habits, they might have problems of overweight in the future; this was seen more frequently in women than in men. Young adults had the highest percentage of self-perception of future overweight with current habits for both men and women. Most of the participants (325/452, 71.9%) would use the chatbot for obesity risk assessment and recommended it (406/452, 89.8%), mostly by talking about it, followed by through the medical centers and in their social networks. In addition, most participants believed that it would help to prevent obesity. They would prefer functionality regarding physical activity and diet recommendations, as well as about obesity risk assessment. Participants preferred colors in the field of obesity and overweight were, in order from highest to lowest, green, blue, and white. Participant's graphical preferences were based on colors, simplicity, and figures. As well, quite a few of them would like a character associated with the app ("Wireframes results" and Figure S2 in Multimedia Appendix 1).

From the expert panel interviews, we identified the personal, diet, and physical activity questions, as well as the status assessment method (Wakastatus), already described in the chatbot's design and functionality section.

Usability Test Results

Participants were volunteer students recruited face to face in the Design School. In total, 61 students (young adults, mean age 20.5 years) participated in the usability test. All participants used a smartphone with Telegram previously installed on it. All participants were able to start Wakamola in Telegram without help, although most of them were not regular users of this messaging system. As a result, most users, when asked, would prefer that Wakamola be a separate app that could be installed on their mobile phone without Telegram. Most participants were able to understand all questions in the Personal, Diet, and Activity sections; however, they considered the Diet section to contain too many questions (23/61, 38%), while the number of questions was acceptable in other sections.

According to the SUS questionnaire [36], about half of the participants indicated acceptable usability. Further information about usability results is in Multimedia Appendix 1.

Pilot Study Results

We carried out a pilot study with 85 university students recruited face to face. We filtered participants that completed all sections, 74 people in total (54 female, 20 male), for the data analysis. The mean age was 20.7 years, and the mean weight was 62.65 kg (SD 10.21). There were no participants with obesity-related diseases such as hypertension, diabetes, high cholesterol, or cardiovascular disease. The participants were from 55 different living areas according to their zip codes, most of them near the university area.

The percentage of people with overweight was 6.8% (5/74 people), while the percentage of people with underweight was higher at 10.8% (8/74 people). No obesity cases were detected in the sample.

The mean BMI was 21.4 (SD 2.41), which corresponds with normal weight. The mean Wakastatus was 78.3 (SD 10.67) on a scale of 1 to 100, mean Diet score was 63.6 (SD 4.67), mean Activity score was 65.3 (SD 32.91), and mean social network score was 26.6 (SD 13.12).

The most consumed types of food were olive oil, milk and derivatives, cereals, vegetables, and fruits. The less consumed types of food were seafood, butter, French fries, and sweetmeats. The consumption of alcohol and soft drinks was also low (Table 2).

Participants practiced physical activity regularly during the week. They spent a mean of 30.57 hours per week sitting and 7.02 hours per day resting. Table 3 shows the sample physical activity, sitting, and sleep hour habits in a week.

Table 2. Types of foods consumed weekly.

Food type	Units per week, mean
Seafood	0.54
Soft drinks with sugar	0.62
Butter	0.67
Alcohol drinks	0.78
French fries	1.18
Sweetmeats	1.22
Blue fish	1.64
Rice	2.04
Legumes	2.04
White fish	2.07
Sausage	2.13
Meats	2.27
Other oils	2.36
Cheeses	2.43
Nuts	2.65
Fruits	2.83
Vegetables	3.17
Cereals and derivatives	3.17
Milk and derivatives	4.43
Olive oil	12.72

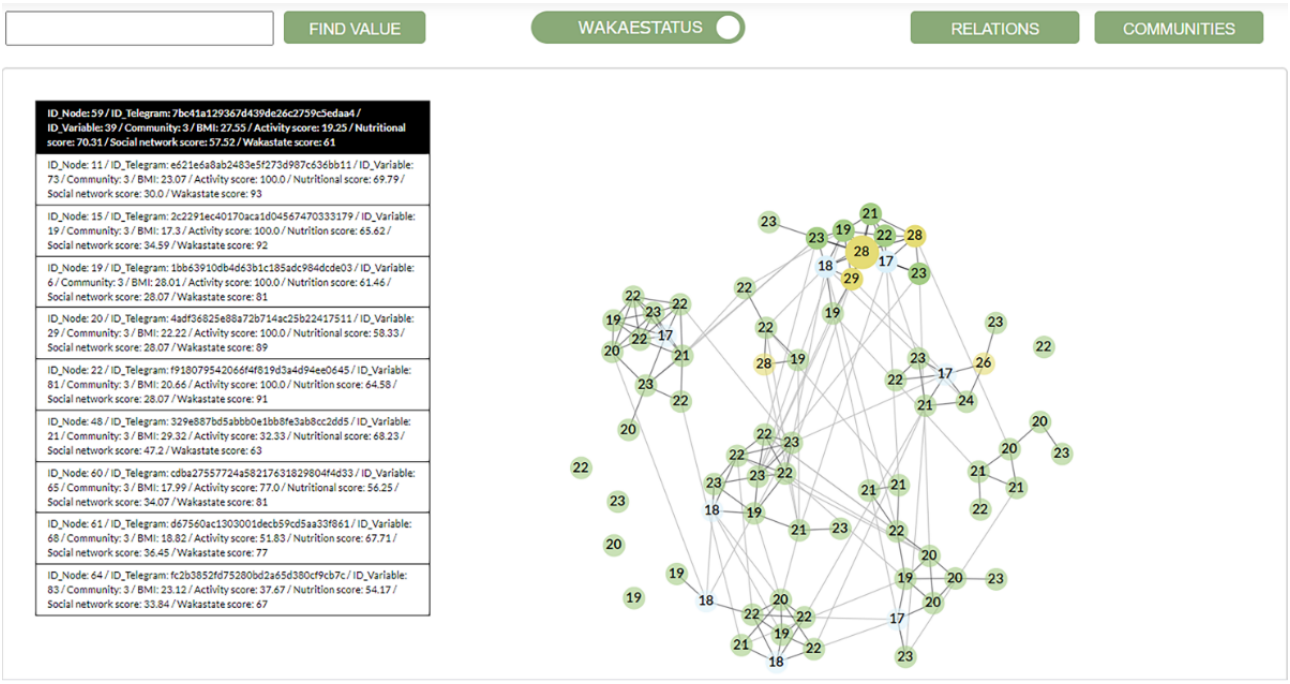
Table 3. Physical activity and sleep hours in mean values per week.

Activity	Values, mean
Vigorous physical activities (times per week)	2.34
Vigorous physical activities (minutes)	33.97
Moderate physical activities (times per week)	5.11
Moderate physical activities (minutes)	35.76
Walked at least 10 continuous minutes (days per week)	5.80
Walking time (minutes)	34.26
Sitting (hours per week)	30.57
Sleep (hours per day)	7.02

We applied the online tool [74] to interpret collected data and showed it as a network graph with 74 users and 178 relations, including 5 users without relations and 12 social groups (communities) (Figure 2). Focusing on the 8 communities with more than one member, 3 of them (38%) had members with overweight, and 6 had members with underweight (75%). All individuals without connections were normal weight. The biggest community had 12 members; it was also the community with the highest percentage of overweight members, with 3 cases (25%).

Figure 2 shows users as nodes colored and labeled according to their BMI value: blue (underweight, <18.5), green (normal weight, 18.50-24.9), yellow (overweight, ≥25), or red (obesity, ≥30). The Wakastatus option allows it to be shown in the nodes. In Figure 2, a user has been selected and highlighted in a community; a table shows the BMIs and scores of his or her relations and contacts.

Figure 2. Target population network and communities representation, with nodes of the same community linked with dark gray edges and a user selected. BMI is shown inside nodes, and colors are based on BMI: blue (underweight, <18.5), green (normal weight, 18.50-24.9), yellow (overweight, ≥25), red (obesity, ≥30). BMI: body mass index.



Discussion

Summary of Findings

We have translated standard questionnaires, traditionally used to collect data about sociodemographics, diet, and physical activity, to a novelty Telegram's chatbot [73]. As well, we have defined a new user-friendly score to assess the user's obesity risk based on his or her diet, physical activity, BMI, and social contacts' lifestyles. Gamification principles have guided the design of the chatbot to help create a positive user experience. Moreover, we have confirmed that people are concerned about their weight and that they consider mHealth apps to be likely to help obesity prevention, as they are interested in using them. In a pilot study deployed in the academic community, we have been able to create a social network to study social factors influencing obesity. The researchers can access an online tool to graphically show the social network to aid data interpretation.

Survey Findings

From the survey about users' needs, we realized a need regarding overweight and obesity apps. This result could be linked with registered participants' worry about their overweight, as 176 out of 452 (38.9%) indicated self-perception of being overweight, and 217 (48.0%) indicated that they could become overweight in the future with their current habits. Moreover, 325 out of 452 (71.9%) participants would use an app to know their obesity risk. As well, 406 out of 452 (89.8%) of them would recommend it. Furthermore, the number of survey responders (452 people) could be an indicator reflecting the concern about overweight and obesity in the university community involved in this study.

Weight management apps represent a popular area of mHealth today [37]. However, there is a need for trustable overweight and obesity apps; most of the commercial mobile apps for weight loss and management lack important evidence-based features, do not involve health care experts in their development process, and have not undergone rigorous scientific testing [38]. Wakamola's chatbot could contribute to cover this need because it involves experts, is based on scientific evidence, and has been subjected to an exhaustive testing process.

Regarding the 52 wireframes scored, we finally selected one based on a character (Wakamola). This selection allowed us to implement personification, the attribution of a personal nature or human characteristics [76] to the chatbot. Personification has a positive effect on the user experience [39]. The introduction of humanlike cues in a chatbot increase users' emotional connection [18]. As well, previous studies suggest a significant effect of anthropomorphic design features (human characteristics) on perceived usefulness, with a strength 4 times the size of the effect of functional chatbot features [40].

The Chatbot as a Feasible Tool to Collect Data

We here propose a chatbot as a novel tool to collect data associated with overweight and obesity. Chatbots could help to collect data in a longitudinal and long-term way [8] that would be difficult and time-consuming with traditional methods, such as standardized questionnaires [77]. Several studies state that a feature that can engage users in completing questionnaires is their presentation through a chatbot [8,18]. Users are more likely to answer questions through a chatbot than in a questionnaire or interview because they connect them with entertainment and novelty, and they are curious about them [19]. Moreover, there is previous research about the application of chatbots to collect data associated with obesity and overweight [10-13]. Furthermore, the use of chatbots has also extended to other health fields such as oncology [78].

We collected data in a pilot study with 85 people. Analyzing the data obtained from the pilot, we found a percentage of people with overweight of 7% (5/74), while the percentage of people with underweight was higher (8/74, 11%); no obesity cases were detected in the sample. The presence of underweight cases could be explained by the higher representation of women in the sample (54/74, 73%), previous studies indicates that women were more likely to be underweight than men [79].

The most consumed types of food were olive oil, milk and derivatives, cereals, vegetables, and fruits, all of which are types of foods associated with the Mediterranean diet [41]. However, other dietary habits with restricted consumption in the Mediterranean Pyramid [41], such as consumption of butter, French fries, sweetmeats, alcohol, and soft drinks, were low [41]. Moreover, most of the participants practiced regular physical activity (Table 1) and slept for a mean of 7 hours, which is a good rest time in adults [80]. These results could explain the participants' mean BMI of 21.4 (SD 2.41), which corresponds to normal weight.

We applied the developed online tool [74] to interpret collected data and showed it as a network graph with 74 nodes and 178 relations, including 5 nodes without relations, and 12 communities based on the Louvain algorithm [75] (Figure 2). The biggest community had 12 members; it was also the community with the highest percentage of overweight members (3 cases out of 74, 25%). Further research would need to study if there is an overweight "contagion" effect in this community [5] or if individuals with similar BMIs are clustering together into this group [6].

This approach and further development of the tool would support the study of overweight and obesity causes, not only from the point of view of the habits of people, but also from the perspective of the influence of their relationships and socioeconomic environment. We recall that previous studies have used social network analysis to study the overweight and obesity problem [81,82] (Multimedia Appendix 1, "Social networks influence in the development of Obesity and Overweight"). It should be noted that the relationships created in the chatbot also specify if it is a relation with a person from home, a family member, a friend, or a coworker, so further research should approach the influence of these subnetworks on the population under study regarding overweight and obesity.

Chatbots to Assess Lifestyle

In Wakamola, diet is scored based on the type and frequency of foods, and physical activity habits are also scored; these are relevant parameters to control body weight [2]. The user's weight and height are also collected to calculate their BMI, which is a widely applied fat mass indicator parameter. Users are informed about their BMI, which is a value unknown to most people, and warned if it is over the recommended values for a normal weight.

Moreover, users get a global score of their status according to input data (Wakastatus), although this score needs further study, for example, regarding the correlation of BMI with defined Wakastatus, diet, activity, and social scores, as well as with other overweight and obesity indicators.

Lessons Learned About the Wakamola Chatbot Design

People are curious about what chatbots are and how they work as a recent technology, which is reflected by the interest in Wakamola in the media after its launch [83]. Based on our experience, people are more likely to use a chatbot in a messaging platform they already use than to install another app in their phones or visit a website. However, after the first approach to the chatbot, people expect more feedback to engage the app. The obesity risk assessment alone is not enough; people ask for recommendations about diet and physical activity (general and personalized), tracking (diet, physical activity, calories consumed), community, sharing progress, positive messages, information about nutrition and healthy habits, success stories, syncing with activity bracelets, and rewards for improving, among others (Table S4 in Multimedia Appendix 1).

The use of a character with personalization helps users to empathize with it and promote the app's use. Based on our experience with Wakamola so far, we know that people want to meet Wakamola after seeing the character image. However, after starting the chatbot, people expect more feedback to become regular users; most of the users use it one time. Thus, the chatbot needs additional effort to improve engagement to enable long-term control studies.

The usability and acceptance problems detected were mainly related to the dependency on the Telegram platform ("Usability test results" in Multimedia Appendix 1). Participants that were not Telegram users before using Wakamola needed help to share it. Moreover, they were unlikely to use it by their own initiative. The decision to develop the chatbot in a third-party platform has advantages, such as speeding up the development and removing the requirement for regular users to install a new app to use the chatbot, saving storage space in their phones. However, this dependency reduces the acceptance of the app for people unfamiliar with the platform because they think that the effort required is higher than installing only an independent app. Moreover, this dependency slows down the expansion of the app and therefore the creation of the network to support the study of social causes of obesity. Thus, we consider that the chatbot Wakamola needs to be multiplatform and an independent online chatbot to reach the maximum number of potential users. As well, the sharing procedure requires an in-depth study to achieve stickiness and usability. Furthermore, the perception of trust is fundamental for acceptance and to extend the use of the chatbot. People would open an invitation to a chatbot only if it includes information about the app objectives and comes from a reliable source.

Moreover, the number of chatbot messages needs to be limited to avoid user fatigue and abandonment. Thus, the Wakamola Diet section in particular needs to be shortened.

From the data collection perspective, the Wakamola chatbot enables the definition of different instances, which could be useful to perform parallel pilot studies in target populations. Two new pilot studies are in process, involving 1500 people so far [84,85].

Conclusions

The Wakamola chatbot provides a tool to collect linked data about users' sociodemographics, overweight- and obesity-related diseases, diet and physical activity habits, BMI, social network, and environment. All these data could aid the study of overweight and obesity in a target population. Moreover, the social network created with the chatbot allows the study of overweight and obesity from a social approach; an online tool has been developed to support it. As well, the chatbot is an end user tool for self-assessment of overweight and obesity risk. Results indicate that this new chatbot meets the needs of both end users and experts, although usability and feedback should keep improving. Moreover, its user-centered design would contribute to the chatbot's usability and acceptance in real scenarios.

However, we are aware of the limitations of this preliminary study. The cohort in the pilot study might not be representative due to selection bias. We plan to apply Wakamola in wide populations in a real context to analyze the data and social network. Moreover, we intend to study the feasibility of the chatbot to help overweight and obesity screening and interventions.

Further studies will focus on Wakamola's usability improvement, collecting data in large populations for social network analysis, the chatbot's messaging multiplatform compatibility, the study of gamification perception and its effects on the user, and chatbots' performance in comparison to traditional graphical user interfaces in applications in the field of obesity and overweight.

In addition, a Wakastatus score validation is required to clarify its perception by users and its feasibility to assess users' obesity risks.

Acknowledgments

The authors gratefully acknowledge designers María Dolores Blanco, Ángel Esteban, and Marta Lavall for their contribution as graphical designers for the Wakamola chatbot, as well as the support to this research provided by Salvador Tortajada from the Scientific Unit of Business Innovation at the Institute of Corpuscular Physics.

Moreover, the authors acknowledge the funding support for this study provided by the CrowdHealth Project (Collective Wisdom Driving Public Health Policies, 727560).

Finally, the authors thank the subjects whose participation made this study possible.

Authors' Contributions

SAC, VBS, JAC, AF, MGP, JFMT, MRA, SSA, YCL, RVM, and JMGG contributed to the software design; SAC, VBS, MGP, JAC, and JMGG to the software development; SAC, VBS, JAC, and JMGG to the analysis of the results and to the writing of the manuscript; and LFL to the critical review of the manuscript.

Conflicts of Interest

Author LFL is the cofounder and Chief Scientific Officer of Adhera Health Inc (USA). No software from this digital health company has been used in this study, where LFL only had a scientific advisory role. The other authors declare no conflicts of interest.

Multimedia Appendix 1
Wakamola supplementary material.
[DOCX File , 413 KB - medinform_v9i4e17503_app1.docx]

Multimedia Appendix 2
Wakamola chatbot.
[DOCX File , 2159 KB - medinform_v9i4e17503_app2.docx]

References

1. OECD. Obesity and the Economics of Prevention. 2010. URL: https://www.oecd-ilibrary.org/social-issues-migration-health/obesity-and-the-economics-of-prevention_9789264084865-en [accessed 2021-04-07]
2. Lonie S, Farley D, U.S. Department of Health and Human Services and U.S. Department of Agriculture. 2015-2020 Dietary Guidelines for Americans, 8th Edition. 2015 Dec. URL: http://health.gov/dietaryguidelines/2015/guidelines/ [accessed 2021-04-06]
3. WHO. The challenge of obesity - quick statistics. URL: https://www.euro.who.int/en/health-topics/noncommunicable-diseases/obesity/data-and-statistics [accessed 2021-01-14]
4. Hruby A, Hu FB. The Epidemiology of Obesity: A Big Picture. Pharmacoeconomics 2015 Jul;33(7):673-689 [FREE Full text] [doi: 10.1007/s40273-014-0243-x] [Medline: 25471927]

5. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. N Engl J Med 2007 Jul 26;357(4):370-379. [doi: 10.1056/NEJMsa066082] [Medline: 17652652]
6. Bahr DB, Browning RC, Wyatt HR, Hill JO. Exploiting social networks to mitigate the obesity epidemic. Obesity (Silver Spring) 2009 Apr;17(4):723-728 [FREE Full text] [doi: 10.1038/oby.2008.615] [Medline: 19148124]
7. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]
8. Biduski D, Bellei EA, Rodriguez JPM, Zaina LAM, De Marchi ACB. Assessing long-term user experience on a mobile health application through an in-app embedded conversation-based questionnaire. Computers in Human Behavior 2020 Mar;104:106169. [doi: 10.1016/j.chb.2019.106169]
9. Kelli HM, Witbrodt B, Shah A. The future of mobile health applications and devices in cardiovascular health. Euro Med J Innov 2017 Jan;2017:92-97 [FREE Full text] [Medline: 28191545]
10. Holmes S, Moorhead A, Bond R, Zheng H, Coates V, McTear M. WeightMentor: A New Automated Chatbot for Weight Loss Maintenance. 2018 Presented at: 32nd International BCS Human Computer Interaction Conference (HCI); July 4-6, 2018; Belfast, UK. [doi: 10.14236/ewic/hci2018.103]
11. Filler A, Farpour-Lambert N, Barata F, L'Allemand D, Gindrat P, Volland D, et al. Text-based Healthcare Chatbots Supporting Patient and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity. 2017 Presented at: International Conference on Design Science Research in Information System and Technology (DESRIST) 2017; May 30-June 1, 2017; Karlsruhe, Germany. [doi: 10.1007/978-3-319-59144-5_36]
12. Huang A, Yang M, Huang C, Chen Y, Wu M, Chen K. A Chatbot-supported Smart Wireless Interactive Healthcare System for Weight Control and Health Promotion. 2018 Presented at: 2018 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM); 2018; Bangkok, Thailand p. 1791-1795. [doi: 10.1109/ieem.2018.8607399]
13. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med 2019 May 16;9(3):440-447 [FREE Full text] [doi: 10.1093/tbm/ibz043] [Medline: 31094445]
14. Rodríguez-Pérez C, Molina-Montes E, Verardo V, Artacho R, García-Villanova B, Guerra-Hernández EJ, et al. Changes in Dietary Behaviours during the COVID-19 Outbreak Confinement in the Spanish COVIDiet Study. Nutrients 2020 Jun 10;12(6):1730 [FREE Full text] [doi: 10.3390/nu12061730] [Medline: 32531892]
15. Haddad C, Zakhour M, Bou Kheir M, Haddad R, Al Hachach M, Sacre H, et al. Association between eating behavior and quarantine/confinement stressors during the coronavirus disease 2019 outbreak. J Eat Disord 2020;8:40 [FREE Full text] [doi: 10.1186/s40337-020-00317-0] [Medline: 32879730]
16. Reyes-Olavarría D, Latorre-Román PÁ, Guzmán-Guzmán IP, Jerez-Mayorga D, Caamaño-Navarrete F, Delgado-Floody P. Positive and Negative Changes in Food Habits, Physical Activity Patterns, and Weight Status during COVID-19 Confinement: Associated Factors in the Chilean Population. Int J Environ Res Public Health 2020 Jul 28;17(15):A [FREE Full text] [doi: 10.3390/ijerph17155431] [Medline: 32731509]
17. Sánchez-Sánchez E, Ramírez-Vargas G, Avellaneda-López Y, Orellana-Pecino JI, García-Marín E, Díaz-Jimenez J. Eating Habits and Physical Activity of the Spanish Population during the COVID-19 Pandemic Period. Nutrients 2020 Sep 15;12(9):2826 [FREE Full text] [doi: 10.3390/nu12092826] [Medline: 32942695]
18. Go E, Sundar SS. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. Computers in Human Behavior 2019 Aug;97:304-316 [FREE Full text] [doi: 10.1016/j.chb.2019.01.020]
19. Brandtzaeg B, Følstad A. Why People Use Chatbots. : Springer International Publishing, Cham; 2017 Presented at: International Conference on Internet Science (INSCI) 2017; November 22-24, 2017; Thessaloniki, Greece p. 377-392. [doi: 10.1007/978-3-319-70284-1_30]
20. Johnson D, Deterding S, Kuhn KA, Staneva A, Stoyanov S, Hides L. Gamification for health and wellbeing: A systematic review of the literature. Internet Interv 2016 Nov;6:89-106 [FREE Full text] [doi: 10.1016/j.invent.2016.10.002] [Medline: 30135818]
21. Cechetti NP, Bellei EA, Biduski D, Rodriguez JPM, Roman MK, De Marchi ACB. Developing and implementing a gamification method to improve user engagement: A case study with an m-Health application for hypertension monitoring. Telematics and Informatics 2019 Aug;41:126-138. [doi: 10.1016/j.tele.2019.04.007]
22. Harms J, Biegler S, Wimmer C, Kappel K, Grechenig T. Gamification of Online Surveys: Design Process, Case Study, and Evaluation. 2015 Presented at: IFIP Conference on Human-Computer Interaction (INTERACT) 2015; September 14-18, 2015; Bamberg, Germany. [doi: 10.1007/978-3-319-22701-6_16]
23. Hamari J. Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. Electronic Commerce Research and Applications 2013 Jul;12(4):236-245. [doi: 10.1016/j.elerap.2013.01.004]
24. Sardi L, Idri A, Fernández-Alemán JL. A systematic review of gamification in e-Health. J Biomed Inform 2017 Jul;71:31-48 [FREE Full text] [doi: 10.1016/j.jbi.2017.05.011] [Medline: 28536062]
25. Portela M, Granell-Canut C. A New Friend in Our Smartphone?: Observing Interactions with Chatbots in the Search of Emotional Engagement. 2017 Presented at: XVIII International Conference on Human Computer Interaction; 2017; New York, NY p. 48-48. [doi: 10.1145/3123818.3123826]

26. Georgsson M, Staggers N, Årsand E, Kushniruk A. Employing a user-centered cognitive walkthrough to evaluate a mHealth diabetes self-management application: A case study and beginning method validation. J Biomed Inform 2019 Mar;91:103110 [FREE Full text] [doi: 10.1016/j.jbi.2019.103110] [Medline: 30721757]
27. Siutila M. The gamification of gaming streams. 2018 Presented at: GamiFIN Conference 2018; May 21-23, 2018; Pori, Finland URL: http://ceur-ws.org/Vol-2186/paper16.pdf
28. Puhl RM, Heuer CA. Obesity stigma: important considerations for public health. Am J Public Health 2010 Jun;100(6):1019-1028. [doi: 10.2105/AJPH.2009.159491] [Medline: 20075322]
29. DeepBot. URL: https://deepbot.deep.sg/ [accessed 2021-01-14]
30. PhantomBot. URL: https://phantombot.github.io/PhantomBot/ [accessed 2021-04-06]
31. Heller E. Psicología del color: Cómo actúan los colores sobre los sentimientos y la razón (Spanish Edition). Barcelona: Gustavo Gili, D.L; 2004:288.
32. Bazán B. La conexión emocional con el color.: Los colores que más y menos gustan en España y sus significados. Revista Sonda: Investigación y Docencia en Artes y Letras 2018;7:275-290 [FREE Full text]
33. Rodríguez IT, Ballart JF, Pastor GC, Jordà EB, Val VA. [Validation of a short questionnaire on frequency of dietary intake: reproducibility and validity]. Nutr Hosp 2008;23(3):242-252. [Medline: 18560701]
34. Van Rossum G, Drake Jr FL. Python tutorial. Amsterdam: Centrum voor Wiskunde en Informatica; 1995.
35. Fadhil A, Schiavo G, Wang Y, Yilma BA. The Effect of Emojis when Interacting with Conversational Interface Assisted Health Coaching System. 2018 Presented at: 12th European Alliance for Innovation International Conference on Pervasive Computing Technologies for Healthcare; 2018; New York, NY p. 378-383. [doi: 10.1145/3240925.3240965]
36. Brooke J. SUS-A quick and dirty usability scale. In: Jordan PJ, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation In Industry. London: CRC Press; 1996:6.
37. Gabrielli S, Dianti M, Maimone R, Betta M, Filippi L, Ghezzi M, et al. Design of a Mobile App for Nutrition Education (TreC-LifeStyle) and Formative Evaluation With Families of Overweight Children. JMIR Mhealth Uhealth 2017 Apr 13;5(4):e48 [FREE Full text] [doi: 10.2196/mhealth.7080] [Medline: 28408361]
38. Rivera J, McPherson AC, Hamilton J, Birken C, Coons M, Peters M, et al. User-Centered Design of a Mobile App for Weight and Health Management in Adolescents With Complex Health Needs: Qualitative Study. JMIR Form Res 2018 Apr 04;2(1):e7 [FREE Full text] [doi: 10.2196/formative.8248] [Medline: 30684409]
39. Smestad TL. Personality matters! Improving the user experience of chatbot interfaces-personality provides a stable pattern to guide the design and behaviour of conversational agents.: NTNU; 2018. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2502575/18507_FULLTEXT.pdf [accessed 2021-04-07]
40. Rietz T, Benke I, Maedche A. The Impact of Anthropomorphic and Functional Chatbot Design Features in Enterprise Collaboration Systems on User Acceptance. 2019 Presented at: 14th International Conference on Wirtschaftsinformatik; February 24-27, 201; Siegen, Germany.
41. Bach-Faig A, Berry EM, Lairon D, Reguant J, Trichopoulou A, Dernini S, Mediterranean Diet Foundation Expert Group. Mediterranean diet pyramid today. Science and cultural updates. Public Health Nutr 2011 Dec;14(12A):2274-2284. [doi: 10.1017/S1368980011002515] [Medline: 22166184]
42. Wasserman S, Faust K. Social network analysis: Methods and applications. Cambridge, UK: Cambridge University Press; 1994.
43. Nam S, Redeker N, Whittemore R. Social networks and future direction for obesity research: A scoping review. Nurs Outlook 2015;63(3):299-317 [FREE Full text] [doi: 10.1016/j.outlook.2014.11.001] [Medline: 25982770]
44. Pachucki MC, Goodman E. Social Relationships and Obesity: Benefits of Incorporating a Lifecourse Perspective. Curr Obes Rep 2015 Jun;4(2):217-223 [FREE Full text] [doi: 10.1007/s13679-015-0145-z] [Medline: 26213644]
45. Powell K, Wilcox J, Clonan A, Bissell P, Preston L, Peacock M, et al. The role of social networks in the development of overweight and obesity among adults: a scoping review. BMC Public Health 2015 Sep 30;15:996 [FREE Full text] [doi: 10.1186/s12889-015-2314-0] [Medline: 26423051]
46. Marko-Holguin M, Cordel SL, Van Voorhees BW, Fogel J, Sykes E, Fitzgibbon M, et al. A Two-Way Interactive Text Messaging Application for Low-Income Patients with Chronic Medical Conditions: Design-Thinking Development Approach. JMIR Mhealth Uhealth 2019 May 01;7(5):e11833 [FREE Full text] [doi: 10.2196/11833] [Medline: 31042152]
47. Steele Gray C, Irfan Khan A, McKillop I, Sharpe S, Cott C. User-centred co-design with multiple user groups: The case of the electronic Patient Reported Outcome (ePRO) mobile application and portal. Int J Integr Care 2019 Aug 08;19(4):439 [FREE Full text] [doi: 10.5334/ijic.s3439]
48. Birnie KA, Campbell F, Nguyen C, Lalloo C, Tsimicalis A, Matava C, et al. iCanCope PostOp: User-Centered Design of a Smartphone-Based App for Self-Management of Postoperative Pain in Children and Adolescents. JMIR Form Res 2019 Apr 22;3(2):e12028 [FREE Full text] [doi: 10.2196/12028] [Medline: 31008704]
49. Tsai CC, Lee G, Raab F, Norman GJ, Sohn T, Griswold WG, et al. Usability and Feasibility of PmEB: A Mobile Phone Application for Monitoring Real Time Caloric Balance. Mobile Netw Appl 2007 Jul 15;12(2-3):173-184. [doi: 10.1007/s11036-007-0014-4]

50. Curtis KE, Lahiri S, Brown KE. Targeting Parents for Childhood Weight Management: Development of a Theory-Driven and User-Centered Healthy Eating App. JMIR Mhealth Uhealth 2015 Jun 18;3(2):e69 [FREE Full text] [doi: 10.2196/mhealth.3857] [Medline: 26088692]
51. Fedele D, Lucero R, Janicke D, Abu-Hasan M, McQuaid E, Moon J, et al. Protocol for the Development of a Behavioral Family Lifestyle Intervention Supported by Mobile Health to Improve Weight Self-Management in Children With Asthma and Obesity. JMIR Res Protoc 2019 Jun 24;8(6):e13549 [FREE Full text] [doi: 10.2196/13549] [Medline: 31237240]
52. Bardus M, Ali A, Demachkieh F, Hamadeh G. Assessing the Quality of Mobile Phone Apps for Weight Management: User-Centered Study With Employees From a Lebanese University. JMIR Mhealth Uhealth 2019 Jan 23;7(1):e9836 [FREE Full text] [doi: 10.2196/mhealth.9836] [Medline: 30672742]
53. Griffin L, Lee D, Jaisle A, Carek P, George T, Laber E, et al. Creating an mHealth App for Colorectal Cancer Screening: User-Centered Design Approach. JMIR Hum Factors 2019 May 08;6(2):e12700 [FREE Full text] [doi: 10.2196/12700] [Medline: 31066688]
54. Dopp AR, Parisi KE, Munson SA, Lyon AR. A glossary of user-centered design strategies for implementation experts. Transl Behav Med 2019 Nov 25;9(6):1057-1064. [doi: 10.1093/tbm/iby119] [Medline: 30535343]
55. Wray TB, Kahler CW, Simpanen EM, Operario D. User-centered, interaction design research approaches to inform the development of health risk behavior intervention technologies. Internet Interv 2019 Mar;15:1-9 [FREE Full text] [doi: 10.1016/j.invent.2018.10.002] [Medline: 30425932]
56. Jefferson UT, Zachary I, Majee W. Employing a User-Centered Design to Engage Mothers in the Development of a mHealth Breastfeeding Application. Comput Inform Nurs 2019 Oct;37(10):522-531. [doi: 10.1097/CIN.0000000000000549] [Medline: 31414995]
57. Zhou L, DeAlmeida D, Parmanto B. Applying a User-Centered Approach to Building a Mobile Personal Health Record App: Development and Usability Study. JMIR Mhealth Uhealth 2019 Jul 05;7(7):e13194 [FREE Full text] [doi: 10.2196/13194] [Medline: 31278732]
58. Lewis JR. IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 1995 Jan;7(1):57-78. [doi: 10.1080/10447319509526110]
59. Pricilla C, Lestari DP, Dharma D. Designing Interaction for Chatbot-Based Conversational Commerce with User-Centered Design. 2018 Presented at: 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA); 2018; Krabi, Thailand. [doi: 10.1109/icaicta.2018.8541320]
60. Vilardaga R, Rizo J, Zeng E, Kientz JA, Ries R, Otis C, et al. User-Centered Design of Learn to Quit, a Smoking Cessation Smartphone App for People With Serious Mental Illness. JMIR Serious Games 2018 Jan 16;6(1):e2 [FREE Full text] [doi: 10.2196/games.8881] [Medline: 29339346]
61. LeRouge C, Ma J, Sneha S, Tolle K. User profiles and personas in the design and development of consumer health technologies. Int J Med Inform 2013 Nov;82(11):e251-e268. [doi: 10.1016/j.ijmedinf.2011.03.006] [Medline: 21481635]
62. Hsieh WT, Su YC, Han HL, Huang MY. A Novel mHealth Approach for a Patient-Centered Medication and Health Management System in Taiwan: Pilot Study. JMIR Mhealth Uhealth 2018 Jul 03;6(7):e154 [FREE Full text] [doi: 10.2196/mhealth.9987] [Medline: 29970356]
63. Morita PP, Yeung MS, Ferrone M, Taite AK, Madeley C, Stevens Lavigne A, et al. A Patient-Centered Mobile Health System That Supports Asthma Self-Management (breathe): Design, Development, and Utilization. JMIR Mhealth Uhealth 2019 Jan 28;7(1):e10956 [FREE Full text] [doi: 10.2196/10956] [Medline: 30688654]
64. Crawford SY, Boyd AD, Nayak AK, Venepalli NK, Cuellar S, Wirth SM, et al. Patient-centered design in developing a mobile application for oral anticancer medications. J Am Pharm Assoc (2003) 2019;59(2S):S86-S95.e1. [doi: 10.1016/j.japh.2018.12.014] [Medline: 30745188]
65. Reis CI, Freire CS, Fernández J, Monguet JM. Patient Centered Design: Challenges and Lessons Learned from Working with Health Professionals and Schizophrenic Patients in e-Therapy Contexts. 2011 Presented at: International Conference on ENTERprise Information Systems (CENTERIS) 2011; October 5-7, 2011; Vilamoura, Portugal. [doi: 10.1007/978-3-642-24352-3_1]
66. Rivest R. The MD5 Message-Digest Algorithm. 1992 Apr. URL: https://www.rfc-editor.org/rfc/pdfrfc/rfc1321.txt.pdf [accessed 2021-04-07]
67. Koivisto J, Hamari J. The rise of motivational information systems: A review of gamification research. International Journal of Information Management 2019 Apr;45:191-210 [FREE Full text] [doi: 10.1016/j.ijinfomgt.2018.10.013]
68. Schmidt-Kraepelin M, Thiebes S, Stepanovic S, Mettler T, Sunyaev A. Gamification in Health Behavior Change Support Systems - A Synthesis of Unintended Side Effects. 2019 Presented at: 14th International Conference on Wirtschaftsinformatik; February 24-27, 2019; Siegen, Germany.
69. Bangor A, Kortum J, Miller J. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. J Usability Stud 2009;4:114-123.
70. Norte Navarro AI, Ortiz Moncada R. [Spanish diet quality according to the healthy eating index]. Nutr Hosp 2011;26(2):330-336. [doi: 10.1590/S0212-16112011000200014] [Medline: 21666971]

71. Mantilla Toloza S, Gómez-Conesa A. El Cuestionario Internacional de Actividad Física. Un instrumento adecuado en el seguimiento de la actividad física poblacional. Revista Iberoamericana de Fisioterapia y Kinesiología 2007 Jan;10(1):48-52. [doi: 10.1016/s1138-6045(07)73665-1]
72. NHLBI Expert Panel on the Identification, Evaluation, Treatment of Overweight Obesity in Adults, National Heart, Lung, and Blood Institute. The Practical guide: identification, evaluation, and treatment of overweight and obesity in adults. 2002. URL: http://hdl.handle.net/2027/umn.31951p00598699p [accessed 2021-04-07]
73. Wakamola Chatbot in Telegram. URL: https://t.me/wakamolabot [accessed 2021-04-06]
74. Asensio-Cuesta S, Blanes-Selva V, García-Gómez JM, Conejero JA, Portolés M. WakaSNA tool. 2019. URL: https://wakamola.webs.upv.es/index.php/herramientas/ [accessed 2021-04-07]
75. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/p10008]
76. Oxford Dictionary, Personification. URL: https://www.lexico.com/definition/personification [accessed 2021-01-14]
77. Hunger JM, Tomiyama AJ. Weight Labeling and Disordered Eating Among Adolescent Girls: Longitudinal Evidence From the National Heart, Lung, and Blood Institute Growth and Health Study. J Adolesc Health 2018 Sep;63(3):360-362 [FREE Full text] [doi: 10.1016/j.jadohealth.2017.12.016] [Medline: 29705495]
78. Bibault JE, Chaix B, Nectoux P, Pienkowsky A, Guillemasse A, Brouard B. Healthcare ex Machina: Are conversational agents ready for prime time in oncology? Clin Transl Radiat Oncol 2019 May;16:55-59 [FREE Full text] [doi: 10.1016/j.ctro.2019.04.002] [Medline: 31008379]
79. Lee Y. Slender women and overweight men: gender differences in the educational gradient in body weight in South Korea. Int J Equity Health 2017 Nov 21;16(1):202 [FREE Full text] [doi: 10.1186/s12939-017-0685-9] [Medline: 29157251]
80. National Sleep Foundation. National Sleep Foundation Recommends New Sleep Times. URL: https://www.sleepfoundation.org/press-release/national-sleep-foundation-recommends-new-sleep-times [accessed 2020-04-15]
81. Marks J, Sanigorski A, Owen B, McGlashan J, Millar L, Nichols M, et al. Networks for prevention in 19 communities at the start of a large-scale community-based obesity prevention initiative. Transl Behav Med 2018 Jul 17;8(4):575-584 [FREE Full text] [doi: 10.1093/tbm/iby026] [Medline: 30016518]
82. Chomutare T. Patient similarity using network structure properties in online communities. 2014 Presented at: IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI); June 1-4, 2014; Valencia, Spain. [doi: 10.1109/bhi.2014.6864487]
83. Wakamola news. URL: https://wakamola.webs.upv.es/index.php/2019/01/28/__trashed/ [accessed 2021-01-14]
84. Wakamola UPV Community. URL: http://sigiloso.itaca.upv.es/wakamolaupv/index.html [accessed 2021-04-07]
85. Wakamola Tavernes Community. URL: http://sigiloso.itaca.upv.es//wakamolatavernes/index.html [accessed 2021-04-06]

Abbreviations

IPAQ: International Physical Activity Questionnaire
mHealth: mobile health
SUS: System Usability Scale
UPV: Universitat Politècnica de València

Edited by C Lovis; submitted 17.12.19; peer-reviewed by E Bellei, S Thiebes; comments to author 26.08.20; revised version received 05.10.20; accepted 20.02.21; published 14.04.21.

Please cite as:
Asensio-Cuesta S, Blanes-Selva V, Conejero JA, Frigola A, Portolés MG, Merino-Torres JF, Rubio Almanza M, Syed-Abdul S, Li YC(, Vilar-Mateo R, Fernandez-Luque L, García-Gómez JM
A User-Centered Chatbot (Wakamola) to Collect Linked Data in Population Networks to Support Studies of Overweight and Obesity Causes: Design and Pilot Study
JMIR Med Inform 2021;9(4):e17503
URL: https://medinform.jmir.org/2021/4/e17503
doi:10.2196/17503
PMID:33851934

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org