JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 9 (2021), Issue 12 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Reviews

Machine Learning Algorithms to Detect Subclinical Keratoconus: Systematic Review (e27363)
Howard Maile, Ji-Peng Li, Daniel Gore, Marcello Leucci, Padraig Mulholland, Scott Hau, Anita Szabo, Ismail Moghul, Konstantinos Balaskas, Kaoru Fujinami, Pirro Hysi, Alice Davidson, Petra Liskova, Alison Hardcastle, Stephen Tuft, Nikolas Pontikos. 3

Artificial Intelligence in Predicting Cardiac Arrest: Scoping Review (e30798)
Asma Alamgir, Osama Mousa, Zubair Shah. 24

Artificial Intelligence–Based Framework for Analyzing Health Care Staff Security Practice: Mapping Review and Simulation Study (e19250)
Prosper Yeng, Livinus Nweke, Bian Yang, Muhammad Ali Fauzi, Einar Snekkenes. 38

Viewpoint

Can Real-time Computer-Aided Detection Systems Diminish the Risk of Postcolonoscopy Colorectal Cancer? (e25328)
Mariusz Madalinski, Roger Prudham. 63

Original Papers

Text Mining of Adverse Events in Clinical Trials: Deep Learning Approach (e28632)
Daphne Chopard, Matthias Treder, Padraig Corcoran, Nagheen Ahmed, Claire Johnson, Monica Busse, Irena Spasic. 66

Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model (e26407)
Hong Wu, Jiatong Ji, Haimei Tian, Yao Chen, Weihong Ge, Haixia Zhang, Feng Yu, Jianjun Zou, Mitsuhiro Nakamura, Jun Liao. 87

Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study (e30970)
Nicolas Paris, Antoine Lamer, Adrien Parrot. 98

Deep Learning–Assisted Burn Wound Diagnosis: Diagnostic Model Development Study (e22798)
Che Chang, Feipei Lai, Mesakh Christian, Yu Chen, Ching Hsu, Yo Chen, Dun Chang, Tyng Roan, Yen Yu. 112

A BERT-Based Generation Model to Transform Medical Texts to SQL Queries for Electronic Medical Records: Model Development and Validation (e32698)
Youcheng Pan, Chenghao Wang, Baotian Hu, Yang Xiang, Xiaolong Wang, Qingcai Chen, Junjie Chen, Jingcheng Du. 128

A Smartphone App (AnSim) With Various Types and Forms of Messages Using the Transtheoretical Model for Cardiac Rehabilitation in Patients With Coronary Artery Disease: Development and Usability Study (e23285)
Jah Choi, Ji Kim, Sunki Lee, Seo-Joon Lee, Seung Shin, Se Park, Eun Park, Woohyeun Kim, Jin Na, Cheol Choi, Seung-Woon Rha, Chang Park, Hong Seo, Jeonghoon Ahn, Hyun-Ghang Jeong, Eung Kim. 142

The Effect of Automated Mammogram Orders Paired With Electronic Invitations to Self-schedule on Mammogram Scheduling Outcomes: Observational Cohort Comparison (e27072)
Frederick North, Elissa Nelson, Rebecca Buss, Rebecca Majerus, Matthew Thompson, Brian Crum. 154

Leveraging National Claims and Hospital Big Data: Cohort Study on a Statin-Drug Interaction Use Case (e29286)
Aurélie Bannay, Mathilde Bories, Pascal Le Corre, Christine Riou, Pierre Lemordant, Pascal Van Hille, Emmanuel Chazard, Xavier Dode, Marc Cuggia, Guillaume Bouzillé. 169

Machine Learning Methodologies for Prediction of Rhythm-Control Strategy in Patients Diagnosed With Atrial Fibrillation: Observational, Retrospective, Case-Control Study (e29225)
Rachel Kim, Steven Simon, Brett Powers, Amneet Sandhu, Jose Sanchez, Ryan Borne, Alexis Tumolo, Matthew Zipse, J West, Ryan Aleong, Wendy Tzou, Michael Rosenberg. 185

Prediction Algorithms for Blood Pressure Based on Pulse Wave Velocity Using Health Checkup Data in Healthy Korean Men: Algorithm Development and Validation (e29212)
Dohyun Park, Soo Cho, Kyunga Kim, Hyunki Woo, Jee Kim, Jin-Young Lee, Janghyun Koh, JeanHyoung Lee, Jong Choi, Dong Chang, Yoon-Ho Choi, Ji Chung, Won Cha, Ok Jeong, Se Jekal, Mira Kang. 200

On Missingness Features in Machine Learning Models for Critical Care: Observational Study (e25022)
Janmajay Singh, Masahiro Sato, Tomoko Ohkuma. 211

Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study (e27386)
Qingyu Chen, Alex Rankine, Yifan Peng, Elaheh Aghaarabi, Zhiyong Lu. 225

Differential Biases and Variabilities of Deep Learning–Based Artificial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study (e33049)
Dongchul Cha, Chongwon Pae, Se Lee, Gina Na, Young Hur, Ho Lee, A Cho, Young Cho, Sang Han, Sung Kim, Jae Choi, Hae-Jeong Park. .
2 3 7

Review

Machine Learning Algorithms to Detect Subclinical Keratoconus: Systematic Review

Howard Maile[1*], MSc; Ji-Peng Olivia Li[2*], MA, FRCOphth; Daniel Gore[2], MD, FRCOphth; Marcello Leucci[2], BA; Padraig Mulholland[1,2,3], PhD; Scott Hau[2], MSc; Anita Szabo[1], MSci; Ismail Moghul[2], PhD; Konstantinos Balaskas[2], BS, MD; Kaoru Fujinami[1,2,4,5], MD, PhD; Pirro Hysi[6,7], MD, PhD; Alice Davidson[1], PhD; Petra Liskova[8,9], MD, PhD; Alison Hardcastle[1], PhD; Stephen Tuft[1,2], MD, FRCOphth; Nikolas Pontikos[1,2], PhD

[1]UCL Institute of Ophthalmology, University College London, London, United Kingdom

[2]Moorfields Eye Hospital, London, United Kingdom

[3]Centre for Optometry & Vision Science, Biomedical Sciences Research Institute, Ulster University, Coleraine, United Kingdom

[4]Laboratory of Visual Physiology, Division of Vision Research, National Institute of Sensory Organs, National Hospital Organization Tokyo Medical Center, Tokyo, Japan

[5]Department of Ophthalmology, Keio University School of Medicine, Tokyo, Japan

[6]Section of Ophthalmology, School of Life Course Sciences, King's College London, London, United Kingdom

[7]Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom

[8]Department of Paediatrics and Inherited Metabolic Disorders, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic

[9]Department of Ophthalmology, First Faculty of Medicine, Charles University and General University Hospital, Prague, Czech Republic

[*]these authors contributed equally

Corresponding Author:
Nikolas Pontikos, PhD
UCL Institute of Ophthalmology
University College London
11-43 Bath Street
London, EC1V 9EL
United Kingdom
Phone: 44 (0)207608 ext 6800
Email: n.pontikos@ucl.ac.uk

Abstract

Background: Keratoconus is a disorder characterized by progressive thinning and distortion of the cornea. If detected at an early stage, corneal collagen cross-linking can prevent disease progression and further visual loss. Although advanced forms are easily detected, reliable identification of subclinical disease can be problematic. Several different machine learning algorithms have been used to improve the detection of subclinical keratoconus based on the analysis of multiple types of clinical measures, such as corneal imaging, aberrometry, or biomechanical measurements.

Objective: The aim of this study is to survey and critically evaluate the literature on the algorithmic detection of subclinical keratoconus and equivalent definitions.

Methods: For this systematic review, we performed a structured search of the following databases: MEDLINE, Embase, and Web of Science and Cochrane Library from January 1, 2010, to October 31, 2020. We included all full-text studies that have used algorithms for the detection of subclinical keratoconus and excluded studies that did not perform validation. This systematic review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendations.

Results: We compared the measured parameters and the design of the machine learning algorithms reported in 26 papers that met the inclusion criteria. All salient information required for detailed comparison, including diagnostic criteria, demographic data, sample size, acquisition system, validation details, parameter inputs, machine learning algorithm, and key results are reported in this study.

Conclusions: Machine learning has the potential to improve the detection of subclinical keratoconus or early keratoconus in routine ophthalmic practice. Currently, there is no consensus regarding the corneal parameters that should be included for

assessment and the optimal design for the machine learning algorithm. We have identified avenues for further research to improve early detection and stratification of patients for early treatment to prevent disease progression.

(JMIR Med Inform 2021;9(12):e27363) doi:10.2196/27363

KEYWORDS

artificial intelligence; machine learning; cornea; keratoconus; corneal tomography; subclinical; corneal imaging; decision support systems; corneal disease; keratometry

Introduction

Background

Keratoconus is a bilateral ectatic disease of the cornea that can cause visual loss through corneal distortion and scarring [1,2]. The prevalence of keratoconus varies from 1 in 375 people in Northern Europe [3] to as high as 1 in 48 in some ethnic groups [4,5], with studies suggesting a higher incidence in Middle-Eastern, West Indian, and Asian populations with faster progression [6-8]. The onset of the disease typically occurs after puberty, with subsequent progression at a variable rate over 2 to 3 decades [6]. A recent meta-analysis found that patients <17 years are likely to progress more than 1.5 D in K_{max} over 12 months, and those with steeper K_{max} of more than 55 D are likely to have at least 1.5 D K_{max} progression [6].

As the disease advances, corneal distortion can reach a stage where spectacle-corrected vision is inadequate, and patients must rely on soft or rigid contact lenses to achieve good functional vision [9]. However, contact lenses are not always tolerated, and visual impairment can severely affect quality of life [10,11]. In the natural course of the disease, approximately 20% of the patients are offered a corneal transplant to improve their vision but at the risk of postoperative complications (eg, microbial keratitis and inflammation), potential allograft rejection, and transplant failure [7,12,13]. Most individuals with keratoconus are identified because of the symptoms of visual disturbance or an increase in astigmatism at refraction. Therefore, it is inevitable that most individuals with keratoconus are detected at a stage when visual deterioration has already occurred [14].

The detection of keratoconus at an earlier stage has become increasingly relevant since the introduction of corneal collagen cross-linking (CXL). This is a photochemical treatment of the cornea with UV-A light following the application of riboflavin (vitamin B2), which can arrest the progression of keratoconus in 98.3% of the eyes even in relatively advanced cases [15-20]. The benefit of early treatment to minimize visual loss is clear, and there is evidence that it is cost-effective [21-23], but the mechanism to improve early diagnosis by community-based optometrists is challenging because asymptomatic patients with subclinical disease are unlikely to seek review [14]. Improved detection will probably require improved access or efficient community screening with expensive imaging equipment [24].

Machine learning is a branch of artificial intelligence centered on writing a software capable of learning from data in an autonomous fashion by minimizing a loss function or maximizing the likelihood [25]. It can be broadly classified as either supervised or unsupervised learning [26]. In supervised learning, the algorithm is trained with input data labeled with a desired output so that it can predict an output from unlabeled input data [27]. In comparison, in unsupervised learning, the algorithm is not trained using labeled data. Instead, the algorithm is used to identify patterns or clusters in the data [28]. When applied to the field of keratoconus detection, machine learning may be used to analyze a large number of corneal parameters that can be derived from corneal imaging as well as other clinical and biometric measures such as visual acuity and refraction to predict the disease [29]. It can also be applied directly to imaging data to work at the pixel level [30]. Deep learning, a specific branch of machine learning, uses artificial neural networks (NNs) with multiple layers to process input data [31]. It is particularly well suited to the segmentation or classification of corneal images [32]. Both machine learning and deep learning may facilitate superior diagnostic ability that, when implemented as automated screening tools, could result in significant advances in case detection, mitigating both the cost of new imaging hardware and the burden on ophthalmic health care professionals [33]. In addition, through unsupervised learning, it may be possible to discover previously unknown disease subtypes or features [34,35].

Unlike diabetic retinopathy, which uses a widely adopted diagnostic grading system (Early Treatment Diabetic Retinopathy Study) [36] and in which the diagnosis of early disease is based on the presence of discrete entities on the retina (eg, microaneurysms), the diagnostic grading of subclinical keratoconus has not yet reached the same level of consensus [37]. Frequently used grading systems such as Amsler-Krumeich [38] and ABCD [39] do not specifically include a grade for subclinical keratoconus. More detailed information about keratoconus grading systems is available in Multimedia Appendix 1 [37-43].

Case Definition for Keratoconus

Several terms describe the early stage of keratoconus before vision is affected, including forme fruste keratoconus (FFKC), keratoconus suspect, subclinical keratoconus, and preclinical keratoconus. The most commonly used terms are FFKC and subclinical keratoconus, but there is no consensus on their definition [44].

We have included all papers that contain an identifiable subgroup of eyes with any of the aforementioned definitions because of the overlap in the nomenclature and lack of evidence as to which, if any, pose a particular risk for progression to clinical keratoconus. We excluded papers that only consider eyes with established keratoconus.

Objectives

The aim of this study is to critically evaluate the literature on the algorithmic detection of subclinical keratoconus and its equivalent definitions. Advanced keratoconus is relatively easy to diagnose clinically, such that developing machine learning algorithms to identify advanced disease has limited utility. Therefore, we directed this review to publications that have included detection of subclinical keratoconus because identifying these individuals would allow for early treatment with CXL to reduce the likelihood of disease progression and visual loss. We have structured our review both around the different types of available input data (parameters, indices, and corneal imaging systems) and the machine learning algorithms for keratoconus detection. In addition, we investigated the validation methodology within each study and assessed the potential for bias.

Research Questions

Our specific research questions are as follows:

1. Research question 1: What input data types have been used within subclinical keratoconus detection algorithms and how have they performed?
2. Research question 2: What machine learning algorithms have been used for subclinical keratoconus detection and how have they performed?
3. Research question 3: How was algorithm validation handled among the selected manuscripts?

Methods

Search Strategy

We conducted a literature review of the evidence for the utility of machine learning applied to the detection of keratoconus published between January 1, 2010, and October 31, 2020. The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Statement 2009 criteria [45] was followed to search 4 bibliographic databases: MEDLINE, Embase, Web of Science, and Cochrane Library using keyword search on their title, abstract, and keywords. The review was not registered, and no protocol was prepared.

We used the following keyword search for literature review in bibliographic databases: *((keratoconus) OR (cornea* protru*) OR (cornea* ectasia)) AND ((algorithm) OR (machine learn*) OR (deep learn*) OR (artificial intelligence) OR (detect*) OR (diagnos*) OR (screen*) OR (examin*) OR (analys*) OR (investigat*) OR (identif*) OR (discover*) OR (interpret*) OR (test*))*

Inclusion and Exclusion Criteria

We included studies that investigated the detection of early keratoconus or included a subgroup of patients with early disease, as defined by one of the following terms: subclinical keratoconus, FFKC, preclinical keratoconus, suspected keratoconus, unilateral keratoconus (normal fellow eye), and asymmetric ectasia (normal fellow eye) and any definition considered equivalent to the aforementioned terms. The studies should have reported the performance of their model on a data set that was separate from the training data set (often called a validation or a test set). This includes splitting of the data set into training and test sets (eg, 70% training and 30% testing), K-fold cross-validation (an extension of simple splitting, but the process is repeated K times, eg, when K=10, partition the data set into 90% for training the model and 10% for testing, and the process is repeated 10 times by choosing a different 10% partition each time for testing), or evidence of a validation study where the aim is to assess a previously derived model on a new data set (also known as an external validation). Finally, the full-text article should be available, and only papers published in English were considered.

We excluded papers based on the detection of early keratoconus defined as Amsler-Krumeich stages 1 or 2, as this represents established keratoconus with both clinical and topographical features [46].

Data Synthesis

On the basis of the inclusion criteria, 2 reviewers (HM and JPOL) screened the initial results. These results were then screened for the exclusion criteria by HM and NP. The PRISMA diagram is presented in Figure 1. Any disagreements in meeting the inclusion or exclusion criteria were resolved by discussion. Once the set of articles was finalized, 2 reviewers (HM and JPOL) analyzed each article and extracted the following information in a master table presented in Multimedia Appendix 2 [14,47-71]: author and year, title, system, sample source, country, age, gender, number of eyes for each group, diagnosis details, validation details, input details, input types, method, classification groups, sensitivity, specificity, accuracy, precision, area under the receiver operating characteristic curve (AUC), and source code availability. We summarized the most important information for all the results in Table 1. The main effect measures sought were sensitivity and specificity. If these statistics were not directly available from the article, they were calculated manually using their standard definitions [72]. To visually compare the results, we plotted the sensitivity and specificity across all studies for diagnostic criteria and detection systems in Multimedia Appendix 3.

Figure 1. Filtering steps taken to accept or exclude studies in the systematic review.

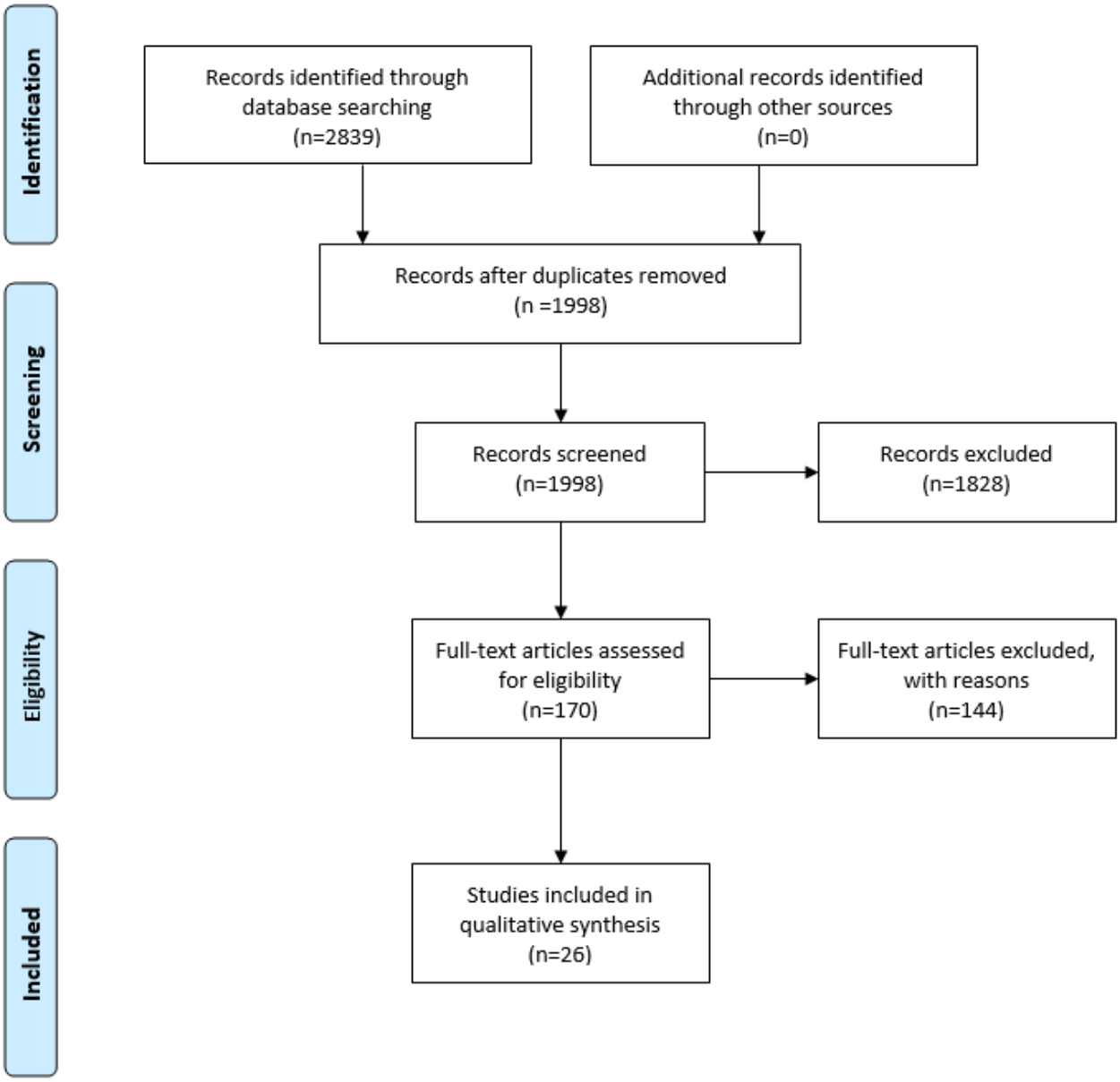


Table 1. Summary of the 26 published studies that included the use of machine learning for the detection of subclinical keratoconus.

Study	System	Number of eyes		Fellow eye[a]	Input types	Method	Results (%)	
		Normal	Subclinical keratoconus				Sensitivity	Specificity
Arbelaez et al [47]	Sirius	1259	426	No	Elevation, keratometry, pachymetry, and aberrometry	SVM[b]	92	97.7
Saad et al [48]	Orbscan	69	34	No	Pachymetry, keratometry, elevation, and Displacement	DA[c]	92	96
Smadja et al [49]	GALILEI	177	47	Yes	Keratometry, pachymetry, elevation, aberrometry, demographic, and indices	DT[d]	93.6	97.2
Ramos-Lopez et al [50]	CSO topography system	50	24	No	Elevation and displacement	Linear regression	33	78
Cao et al [14]	Pentacam	39	49	No	Keratometry, pachymetry, and demographic	RF[e], SVM, K-nearest neighbors, LoR[f], DA, Lasso regression, DT, and NN[g]	94	90
Buhren et al [51]	Orbscan IIz	245	32	No	Keratometry, pachymetry, aberrometry, and elevation	DA	78.1	83.3
Chan et al [52]	Orbscan IIz	104	24	Yes	Pachymetry, keratometry, elevation, and displacement	DA	70.8	98.1
Kovacs et al [53]	Pentacam	60	15	Yes	Keratometry, pachymetry, elevation, indices, and displacement	NN	90	90
Saad et al [54]	OPD-scan	114	62	Yes	Keratometry, aberrometry, and indices	DA	63	82
Ruiz Hidalgo et al [55]	Pentacam HR	194	67	No	Keratometry, pachymetry, and aberrometry	SVM	79.1	97.9
Ruiz Hidalgo et al [56]	Pentacam HR	44	23	No	Keratometry, pachymetry, and indices	SVM	61	75
Xu et al [57]	Pentacam HR	147	77	Yes	Pachymetry, elevation, and keratometry	DA	83.7	84.5
Ambrosio et al [58]	Pentacam+Corvis ST	480	94	Yes	Pachymetry, elevation, keratometry, and Biomechanical	RF, SVM, and LoR	90.4	96
Sideroudi et al [59]	Pentacam	50	55	No	Keratometry	LoR	91.7	100
Francis et al [60]	Corvis ST	253	62	Yes	Biomechanical	LoR	90	91

Study	System	Number of eyes		Fellow eye[a]	Input types	Method	Results (%)	
		Normal	Subclinical keratoconus				Sensitivity	Specificity
Yousefi et al [61]	SS-1000 CASIA	1970	796	No	Elevation, pachymetry, and aberrometry	Unsupervised	88	14
Lopes et al [62]	Pentacam HR	2980	188	Yes	Pachymetry, elevation, indices, and displacement	DA, SVM, naive Bayes, NN, and RF	85.2	96.6
Steinberg et al [63]	Pentacam+Corvis ST	105	50	Yes	Pachymetry, elevation, keratometry, and biomechanical	RF	63	83
Issarti et al [64]	Pentacam	312	90	Yes	Elevation and pachymetry	NN	97.8	95.6
Chandapura et al [65]	RCTVue+Pentacam	221	72	Yes	Keratometry, elevation, pachymetry, aberrometry, and indices	RF	77.2	95.6
Xie at al [66]	Pentacam HR	1368	202	No	Heat maps	CNN[h]	76.5	98.2
Kuo et al [67]	TMS-4+Pentacam+Corvis ST	170	28	No	Heat maps	CNN	28.5	97.2
Shi et al [68]	Pentacam+ultrahigh resolution optical coherence tomography	55	33	Yes	Keratometry, elevation, pachymetry, indices, and demographic	NN	98.5	94.7
Toprak et al [69]	MS-39	66	50	Yes	Keratometry, pachymetry, and displacement	LoR	94	98.5
Issarti et al [70]	Pentacam HR	304	117	Yes	Elevation and Pachymetry	NN	85.2	70
Lavric et al [71]	SS-1000 CASIA	1970	791	No	Keratometry, pachymetry, and aberrometry	25 machine learning methods compares	89.5	96

[a]Fellow eye indicates whether the study defined subclinical keratoconus as the fellow eye of an individual with apparently unilateral keratoconus, with no clinical or topographical features of keratoconus.

[b]SVM: support vector machine.

[c]DA: discriminant analysis.

[d]DT: decision tree.

[e]RF: random forest.

[f]LoR: logistic regression.

[g]NN: neural network.

[h]CNN: convolutional neural network.

Bias Assessment

When assessing bias within the included studies, we used a tailored version of the QUADAS (Quality Assessment of Diagnostic Accuracy Studies)-2 tool [73], which consists of 4 domains: patient selection, index test, reference standard, flow, and timing. The 26 studies were assessed by 3 reviewers (HM, JPOL, and NP) such that each study was assessed by at least 2 reviewers.

Results

Overview

We identified 1998 potentially relevant papers published between 2010 and 2020. After filtering, we included 26 articles in our qualitative analysis. Table 1 summarizes these results, and a more extensive version can be found in Multimedia Appendix 2. To address research question 1, the results are discussed in terms of their input data. Charts displaying

aggregate sensitivity and specificity can be found in Multimedia Appendix 3. To address research question 2, the results are considered in terms of the machine learning algorithms. Figures 2 and 3 present organizational diagrams of data categorization and machine learning algorithms, respectively. To maintain consistency, we opted to use the term *subclinical keratoconus* throughout regardless of the nomenclature used by the original authors. The original term is included in parenthesis, and details of the exact definition can be found in Multimedia Appendix 2.

Figure 2. Organizational diagram of relevant data types reported to be used for the detection of subclinical keratoconus.

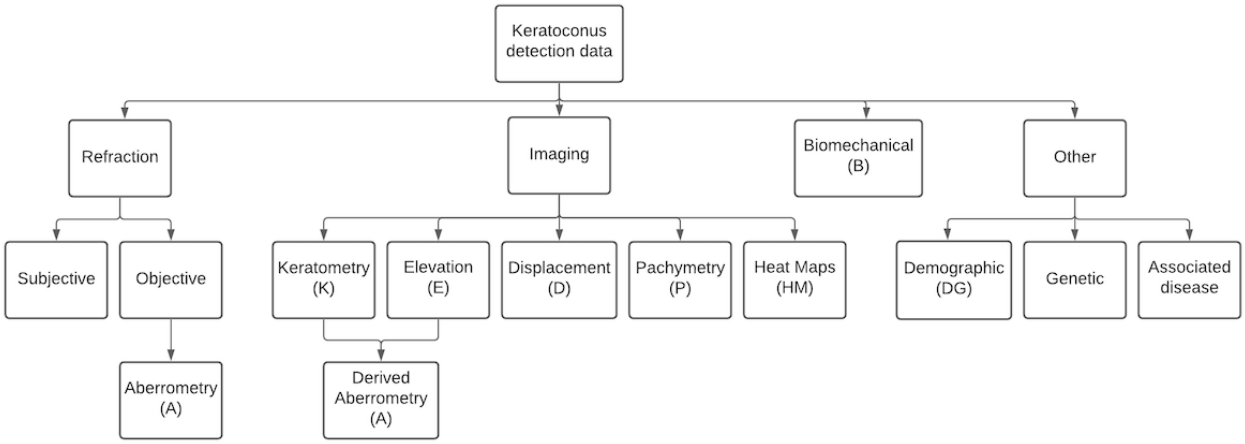
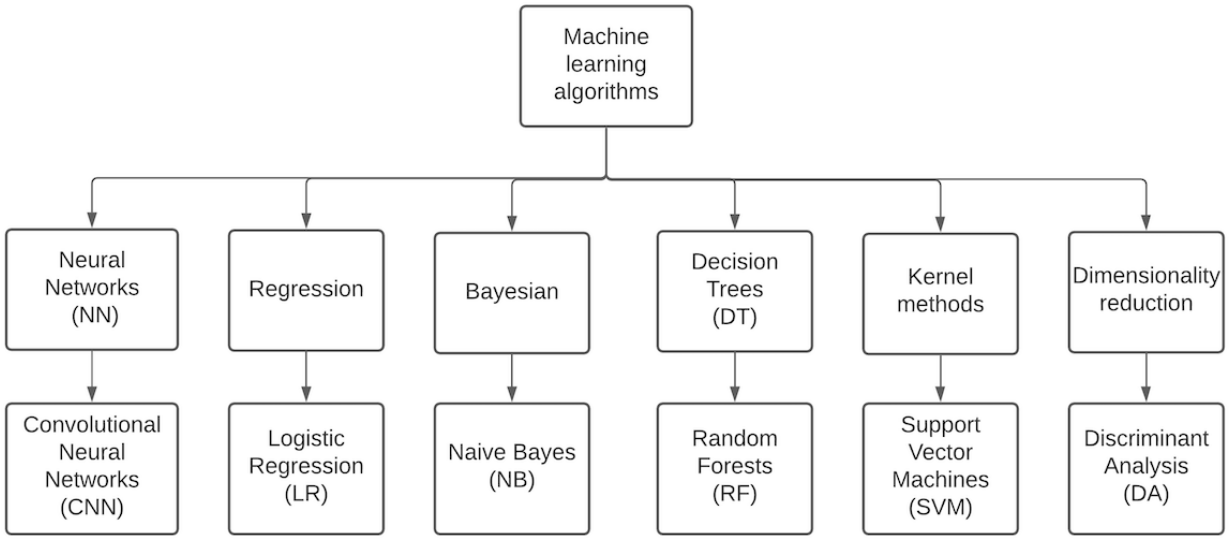


Figure 3. Organizational diagram of relevant machine learning algorithms used for the detection of subclinical keratoconus.



Research Question 1: What Input Data Types Have Been Used Within Subclinical Keratoconus Detection Algorithms and How Have They Performed?

This section is subdivided according to the input data types used for the detection of subclinical keratoconus, as presented in the organizational chart in Figure 2.

Aberrometry

Aberrometry was used to detect subclinical keratoconus in 31% (8/26) of the papers [47-49,51,55,61,65,71]. Aberrations are produced by imperfections in the optical quality of the refracting surface of the eye, including the cornea and the lens. Higher-order aberrations (HOAs) are measured from the distortion of a plane wavefront of light passing through the optics of the eye. However, HOAs can also be derived indirectly from the measurement of any distortion (eg, elevation) of the corneal surfaces. They can be described as a set of Zernike polynomials or with Fourier analysis. Using the Zernike method, aberrations can be subclassified as lower-order aberrations and HOAs. Lower-order aberrations include simple defocus (myopia or hyperopia) and regular astigmatism, which account for approximately 90% of the refractive error of the normal eye [74]. The most clinically relevant HOAs are spherical aberration, coma, and trefoil that cannot be corrected by glasses or a soft contact lens. In keratoconus, the irregular distortion of the front and back surfaces of the cornea causes visually significant HOAs. Arbelaez et al [47] analyzed these parameters in their subclinical keratoconus detection model and included a weighted sum of HOAs (known as the Baiocchi-Calossi-Versaci index) and the root mean square of HOAs. Moreover, 5 other studies also used derived Zernike aberrometry data [48,49,51,65,71].

Corneal Imaging Data and Derived Parameters

Overview

Corneal images were used to detect subclinical keratoconus in 96% (25/26) of the papers. There are various acquisition techniques, including Scheimpflug optics (Pentacam [Oculus GmbH] or Sirius [CSO]), anterior segment optical coherence tomography (AS-OCT; MS-39 [CSO], or CASIA [Tomey]), and horizontal slit-scanning systems such as Orbscan II (Bausch & Lomb). These systems incorporate a software that processes the images to derive numerical indices or secondary images, such as heat maps, to visualize various aspects of corneal shape. These parameters can be classified as measurements of the corneal surface radius of curvature (keratometry), elevation or depression of a point on the corneal surface from the mean (elevation map), corneal thickness (pachymetry), or displacement from the apex of the cornea. Figure 4 illustrates the main parameter types in a schematic diagram. Figure 5 shows an example of the Pentacam heat map for an eye with subclinical keratoconus. See Multimedia Appendix 4 for an example of advanced keratoconus (fellow eye for the same patient).

Figure 4. Schematic diagram illustrating the 4 basic corneal parameters that can be measured using corneal imaging. (A) pachymetry. (B) displacement: distance between the apex of the cornea and the point of minimum thickness. (C) and (D) represent 2 methods of calculating the best-fit sphere (BFS). In (C) the BFS is fitted to both the normal peripheral posterior surface (blue) and the abnormal anterior protrusion of the central posterior surface (green). In (D) the BFS is fitted to only the normal peripheral posterior surface (blue) excluding the abnormal central posterior surface (green), leading to a larger relative elevation than in (C). (E) the smallest radius of curve of the astigmatic corneal surface corresponds to the largest refractive power (Kmax) and the largest radius of curve corresponds to the smallest refractive power (Kmin). CCT: central corneal thickness.

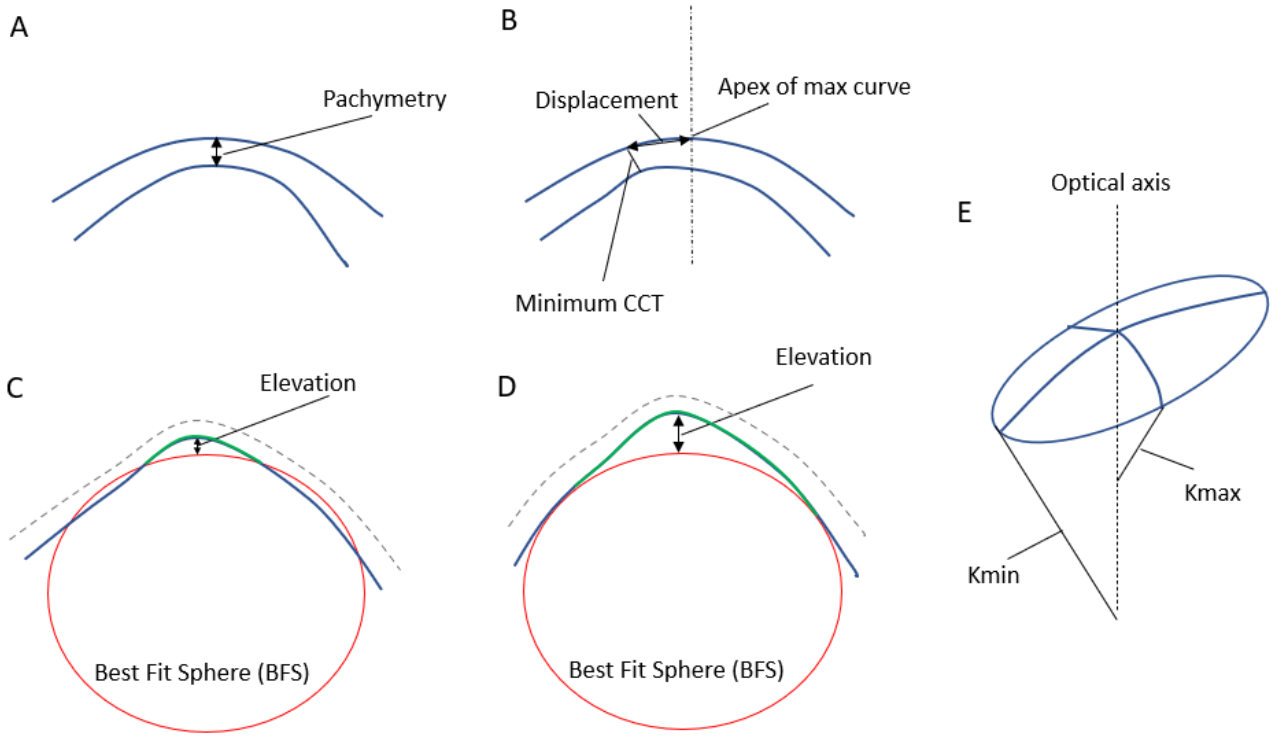
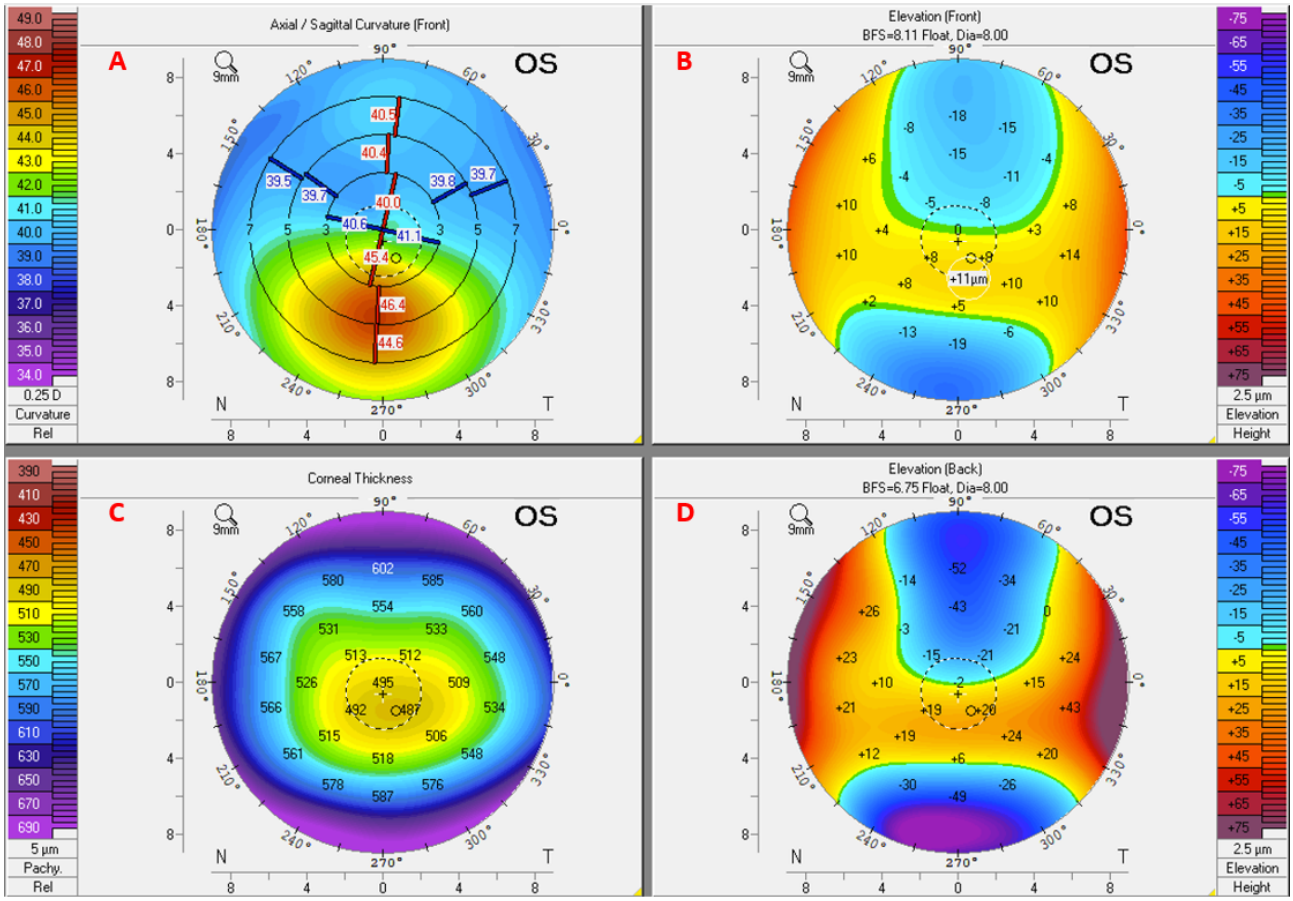


Figure 5. Heat maps of a subclinical keratoconus eye derived from Scheimpflug corneal imaging using the Pentacam HR device. The axial/sagittal map (A) depicts the curvature of the anterior corneal surface in dioptres and shows mild inferior steepening, while the pachymetry map (C) shows thinning in the same region. The front and back elevation maps (B and D, respectively) show a moderate increase in inferior elevation. BFS: best-fit sphere; OS: left eye.



In the following subsections, we briefly discuss the use of quantitative measures derived from corneal imaging when used in isolation or in combination with machine learning models.

Keratometry Parameters

Keratometric data are one of the most commonly used parameters in the literature, with 69% (18/26) of the papers incorporating keratometry as one of the parameters in their model [14,47-49,51-54,56,57,59,61,65,68,71,75]. Keratometric parameters measure the radius of curvature of the anterior or posterior corneal surfaces. Examples include the meridian with the minimum corneal radius of curvature (corresponding to K_{max}) and maximum curvature (corresponding to K_{min}). When looking at individual keratometric parameters derived using Fourier analysis for subclinical keratoconus detection, Sideroudi et al [59] achieved a predictive accuracy of over 90% using higher-order irregularities, asymmetry, and regular astigmatism, primarily in the corneal periphery.

Elevation Parameters

Overall, 62% (16/26) of the papers incorporated elevation parameters in their analysis [47-53,57,58,61-65,68,70]. Elevation represents points above or below the BFS of the corneal surface measured in microns (Figure 4). For the posterior cornea, this is measured either as the divergence from the best fit of the whole posterior corneal diameter or as the divergence from the best fit of the annulus of the peripheral posterior corneal surface outside the central 4 mm [76]. The latter method, the Belin-Ambrosio map, better describes the central corneal elevation, which is a feature of keratoconus. Values can be presented as either color-coded maps or individual parameters such as maximum anterior elevation, maximum posterior elevation, or derived data such as aberrometry.

Posterior corneal curvature consistently outperforms other parameters in the discrimination of subclinical keratoconus [47,49,56,57]. Its inclusion increases the sensitivity of a support vector machine (SVM) from 75.2% to 92% and precision from 57.4% to 78.8% but has a limited impact on specificity [47]. Posterior corneal curvature, measured using a Pentacam (Scheimpflug) device and analyzed using an SVM, was also found to be an important parameter for sensitivity and much less so for specificity and AUC [14]. Similarly, using the Galilei (Scheimpflug) device, the posterior asphericity asymmetry index was found to be the variable with the most discriminatory power when differentiating normal from subclinical keratoconus, followed by corneal volume [49]. Conversely, analysis of anterior surface topographical parameters and aberrometry using the random forest algorithm did not discriminate subclinical keratoconus (very asymmetric ectasia-normal topography) from normal eyes [65].

Saad et al [54] showed that combining parameters obtained from the anterior corneal curvature corneal wavefront and Placido-derived indices lead to a better discriminative ability

between normal and subclinical keratoconus eyes (FFKC) over a Placido-only–based algorithm.

Displacement Parameters

A total of 23% (6/26) of the papers used displacement parameters in their analysis [48,50,52,53,62,75]. These represent measures such as the displacement of the point of minimum corneal thickness from the corneal apex. Of these, 3 papers used the displacement of the thinnest point from the geometric center of the cornea in their model [48,52,77]. Kovacs et al [53] used the vertical and horizontal decentration of the thinnest point and found them to be the best parameters to discriminate normal fellow eyes of keratoconus from control eyes using an NN.

Pachymetry Parameters

Overall, 77% (20/26) of the papers used pachymetry data in their model, making it one of the most commonly used parameters in the literature [14,47-49,51-53,55-58,61-65,69-71]. Pachymetry is the thickness of the cornea, measured using either ultrasound or imaging techniques. Simple examples include central corneal thickness and the thinnest point of the cornea. A reduction in the thickness of the central cornea is a fundamental biomarker of keratoconus [2].

Summary Indices

In total, 23% (6/26) of the papers used summary indices in their model [14,49,53,65,68,78]. In addition to single-parameter measurements (eg, central corneal thickness), tomographic systems such as the Pentacam can combine measurements to compute derived indices that estimate the regularity of corneal shape. Basic indices such as the index of surface variance, index of vertical asymmetry, or index of height asymmetry are formed from multiple data points. Composite indices are formed from other indices and data points. Examples include the Keratoconus index, keratoconus percentage index, and Belin/Ambrosio enhanced ectasia display (BAD-D). In a recent study, Shi et al [68] used 6 indices from the Pentacam along with keratometric, elevation, and pachymetric parameters derived from the Pentacam and ultrahigh resolution optical coherence tomography to create an NN classifier to discriminate between normal and subclinical keratoconus eyes. Using 50 normal eyes, 38 eyes with keratoconus, and 33 eyes with subclinical keratoconus, they achieved 98.5% sensitivity and 94.7% specificity. However, the results require further validation because of the small number of eyes in this group. Furthermore, the authors did not include a comparison between existing detection metrics, such as BAD-D.

Heat Maps

A total of 8% (2/26) of the papers used heat maps in their detection model [66,67]. Modalities such as Scheimpflug and AS-OCT capture images at various corneal meridians and subsequently use these data to derive the heat maps that facilitate visual interpretation of the data, although there is extrapolation of the data in areas between the imaged meridians. For example, the Pentacam can translate the raw images into several types of color heat maps (eg, axial curvature, posterior or anterior elevation, and regional pachymetry) based on the same original tomography data set. Prediction models applied to images often use convolutional NNs (CNNs), and studies applying these methods are discussed in detail in the next section addressing research question 2. To the best of our knowledge, no system has used raw pixel values from Scheimpflug or AS-OCT images directly when detecting subclinical keratoconus.

Biomechanical Data

Overall, 12% (3/26) of the papers incorporated biomechanical data in their analysis [58,60,63]. Corneal biomechanics refers to the distortion response of the cornea to an applied force. The Ocular Response Analyzer (Reichert Ophthalmic Instruments) uses a puff of air directed to the cornea, and the deformation response is measured. Two common indices have been reported: corneal hysteresis and corneal resistance factor. However, there is disagreement regarding their utility in the diagnosis of keratoconus [79,80]. Another device using the same principle is the Corvis ST (Oculus Optikgeräte GmbH), which uses a high-speed Scheimpflug camera to measure distortion in cross-sectional images. Numerous studies have described the application of machine learning to analyze biomechanical data, but very few validated their results; therefore, they have been excluded from this review. Ambrosio et al [58] combined Pentacam and Corvis ST data to create the Tomographic and Biomechanical Index, and this was followed up with a validation study [63]. Francis et al [60] used biomechanical data from the Corvis ST device when diagnosing keratoconus and achieved very high sensitivity (99.5%) and specificity (100%). However, when validating their model, they have only discriminated between 2 groups—a group combining subclinical keratoconus and keratoconus eyes, and a group of normal eyes. This represents an easier problem than including a distinction between normal and subclinical keratoconus eyes.

Demographic Risk Factors

A total of 15% (4/26) of the papers chose to include demographic data, such as age or sex, in their model [14,49,62,68]. Cao et al [14] demonstrated that sex was an important parameter in a minimum set that achieved the highest AUC using the random forest method, although their data set was small (49 subclinical keratoconus and 39 control eyes). Ethnicity, a major association with disease prevalence, risk of progression, disease severity, and acute corneal hydrops in Asian and Black populations [81], was not included in any model, although some studies have examined single ethnicities [66]. Ethnicity as a parameter should be considered by future investigators. No studies included other risk factors such as atopy and eye rubbing as model parameters, and these should be considered in future studies.

Research Question 2: What Machine Learning Algorithms Have Been Used for Subclinical Keratoconus Detection and How Have They Performed?

In most cases, researchers have used combinations of parameters and indices within machine learning algorithms to diagnose subclinical keratoconus. This section is subdivided according to the machine learning techniques that were applied. Figure 3 presents an organizational diagram of the relevant machine learning algorithms. There are several other algorithms, but discussion of these is beyond the scope of the review, and we

have chosen to include only the methods found in our *Results* section.

Neural Networks

Overview

NNs consist of a series of interconnected layers of neurons and are thus loosely modeled on the structure within the human brain. Each neuron computes a nonlinear function of its inputs, and the network is trained until the output aligns optimally with the ground truth labels. Kovacs et al [53] used a combination of 15 keratometric, pachymetric, and elevation parameters in an NN classifier to discriminate healthy corneas from fellow eyes of patients with unilateral keratoconus. The patient data included 60 normal eyes from 30 patients, 60 bilateral keratoconus eyes from 30 patients, and 15 normal eyes from patients presenting with unilateral keratoconus. When classifying the normal eyes of the patients with unilateral keratoconus with clinical grading as a reference, they achieved 90% sensitivity and 90% specificity. They took a novel approach of training on both the eyes of the patients, which allowed them to incorporate the effect of any intereye asymmetry when detecting unilateral keratoconus. Shi et al [68] combined keratometric, elevation, and pachymetric parameters derived from Pentacam images and ultrahigh resolution optical coherence tomography to create an NN classifier for discriminating normal from subclinical keratoconus eyes. Using Pentacam elevation and pachymetry maps within a hybrid NN model, Issarti et al [64] demonstrated superiority over other common diagnostic indices such as BAD-D and topographical keratoconus classification.

Convolutional Neural Networks

When images are used for analysis, NNs with a large number of processing layers such as CNNs are often employed because of their ability to make inferences from 2D or 3D data structures through deep learning [82]. For example, Xie et al [66] used data from 1368 normal eyes, 202 eyes with early keratoconus, 389 eyes with more advanced keratoconus, and 369 eyes with subclinical (suspected) keratoconus to develop an automatic classifier. They achieved 76.5% sensitivity and 98.2% specificity when classifying subclinical keratoconus. However, the heat maps used were produced by Pentacam; therefore, it should be noted that the technique may not be transferable to other systems or even future Pentacam software iterations. Kuo et al [67] included 150 normal, 170 keratoconus, and 28 subclinical eyes in their study and used the Tomey TMS-4 topography system to produce corneal heat maps and trained 3 different CNN architectures (VGG16, InceptionV3, and ResNet152). When attempting to identify the 28 subclinical keratoconus eyes, they applied the VGG16 model and achieved *barely satisfactory* results with an accuracy of 28.5% when a threshold of 50% was applied. These results suggest that subclinical keratoconus cannot yet be detected with high sensitivity using CNNs on heat map images.

Decision Trees

The classification of data in a decision tree uses a binary decision at each node in the tree to determine the branch to take next. Starting from the root, the classification is determined by following each branch to its terminal node. Smadja et al [49] used a decision tree to classify normal, keratoconus, and subclinical (FFKC) keratoconus eyes. They enrolled 177 normal eyes, 148 keratoconus eyes, and 47 subclinical eyes. They used 55 parameters (including curvature, elevation, corneal wavefront, corneal power, pachymetry, and age) collected from the Galilei dual Scheimpflug camera, achieving 93.6% sensitivity and 97.2% specificity when classifying subclinical from normal. Cao et al [14] also evaluated a decision tree algorithm for classifying subclinical keratoconus but achieved lower sensitivity (82%) and specificity (78%). They attributed the comparatively inferior performance to the fact that Smadja et al [49] used additional machine-specific indices that they did not have access to.

Random Forests

Random forests combine a large number of decision trees into a single model [83]. Lopes et al [62] compared this method with other methods (naive Bayes, NNs, SVMs, and discriminant analysis) by training models on 71 post–laser-assisted in situ keratomileusis (LASIK) ectasia eyes, 298 post-LASIK eyes without ectasia, and 183 eyes with keratoconus. They included keratometry, pachymetry, elevation, and various Pentacam indices. The models were validated on an external data set containing 298 normal eyes (stable LASIK), 188 keratoconus eyes (very asymmetric ectasia-ectatic), and 188 subclinical eyes (very asymmetrical ectasia-normal topography). The latter 2 groups were collected from the same set of patients. They found that the random forest model performed best when detecting subclinical eyes with an 85.2% sensitivity. This accuracy is lower than that of other comparable studies, which is probably caused by their inclusion of external validation rather than an inferior model. The authors also note that their model classifies among 3 groups, whereas other related studies (such as that by Arbelaez et al [49]) only classify between 2 groups (eg, subclinical vs normal). This important distinction is expanded upon in the *Discussion* section.

Discriminant Analysis

Discriminant analysis uses a linear combination of variables that optimally separate 2 or more classes of data. Xu et al [57] used this method to classify eyes as either normal, subclinical keratoconus, or keratoconus. In total, 147 normal eyes, 139 eyes with keratoconus, and 77 eyes with subclinical keratoconus were included in the training set and verified on a separate set of 97 normal and 49 subclinical keratoconus eyes. They applied the Zernike fitting method to corneal pachymetry and elevation data derived from the Pentacam and achieved an AUC of 92.8% when discriminating subclinical keratoconus. Saad et al [54] also used discriminant analysis to classify eyes as either subclinical (FFKC) keratoconus or normal. They used a combination of wavefront aberrometry and Placido disc indices in their model with a total of 8 parameters using the OPD-Scan (Nidek Co Ltd). The model was trained on 114 normal and 62 subclinical eyes and validated on 93 normal and 82 subclinical eyes. Using training data only, the model achieved 89% sensitivity and 92% specificity, but when applied to the validation set, the accuracy dropped significantly to 63% sensitivity and 82% specificity. This highlights the need for

external validation when reporting the performance of the detection algorithms.

Support Vector Machines

SVMs translate data into a higher-dimensional space where a dividing line (known as a hyperplane) separates the data such that the distance between the hyperplane and any given data point is maximized [26]. When 8 different machine learning algorithms were compared for classifying subclinical keratoconus on the same data set, SVMs achieved the highest sensitivity (94%) [14]. Arbelaez et al [47] achieved even higher sensitivities using SVMs with a large data set of 1259 normal eyes and 426 with subclinical keratoconus. They used 200 eyes from each group for training and the remainder for testing, achieving 92% sensitivity and 97.7% specificity. Ruiz Hidalgo et al [56] used 25 topographic or tomographic Pentacam-derived parameters to verify their SVM model. They included 131 patients in their study and provided results for 2 classifications from separate hospitals: Antwerp University Hospital and Rothschild Foundation, Paris. When classifying the 4 groups (keratoconus, subclinical, normal, and postrefractive surgery), the sensitivity for subclinical keratoconus detection was 61% compared with that of the Antwerp University Hospital classification and 100% compared with that of the Rothschild classification. This was a comprehensive validation study that compared multiple methods with 2 subjective reference standards. Only a small number of subclinical keratoconus cases (approximately 20) were included in this study, and a larger study is required to verify these results.

Logistic Regression

Logistic regression is commonly used to perform classification from a set of independent variables [26]. It transforms its output using the sigmoid function to return a probability that can then be thresholded for classification. When classifying subclinical keratoconus, 3 studies used this technique exclusively [59,60,75]. Sideroudi et al [59] used logistic regression to explore the diagnostic capacity of Fourier-derived posterior keratometry parameters (spherical component, regular astigmatism, asymmetry, and irregular astigmatism) extracted from Pentacam Scheimpflug images. They included 50 normal eyes, 80 eyes with keratoconus, and 55 with subclinical keratoconus (defined as a clinically normal eye with abnormal topography, where the fellow eye has advanced keratoconus) and validated their model on 30% of the data set. Their model attained 91.7% sensitivity and 100% specificity when classifying between subclinical keratoconus and normal eyes. Although these results are among the best reported, the study has yet to be validated using an external data set. Other studies implemented logistic regression as part of a wider comparison of machine learning algorithms [14,58].

Comparative Studies

Few studies have applied multiple machine learning algorithms to the same data set. Cao et al [14] tested 8 machine learning algorithms on the same data set of 39 normal control eyes and 49 eyes with subclinical keratoconus. Age, sex, and 9 corneal parameters from the Pentacam tomography were used, and the authors found that random forest, SVM, and K-nearest neighbors had the best performance. Random forests had the highest AUC of 0.97, SVM had the highest sensitivity (94%), and K-nearest neighbors had the best specificity (90%). Although they verified their results with 10-fold cross-validation, it would be instructive to repeat the analysis on a larger external data set. Ambrosio et al [58] also performed an analysis across algorithms including logistic regression, SVMs, and random forests to classify between 4 groups: normal, keratoconus, very asymmetrical ectasia-ectatic, and subclinical keratoconus (very asymmetric ectasia-normal). They used both Scheimpflug tomography and biomechanical data and included 480 normal eyes, 204 eyes with keratoconus, 72 eyes classified as very asymmetrical ectasia-ectatic, and 94 subclinical keratoconus eyes. When considering subclinical keratoconus, the random forest model performed the best, with 90.4% sensitivity and 96% specificity. The final model was named the Tomography and Biomechanical Index and was validated by leave-one-out cross-validation, resulting in as many models as there were subjects (N=850). Lopes et al [62] also performed a comparative analysis and found that random forests performed best when trying to classify 3 groups of eyes (including subclinical eyes). Lavric et al [71] provided the largest comparative study for detecting subclinical keratoconus. The authors included 1970 normal eyes, 390 eyes with keratoconus, and 791 subclinical (FFKC) keratoconus eyes in their study and used keratometric, pachymetric, and aberrometric data from the CASIA AS-OCT system in their analysis across 25 different machine learning algorithms. When they classified the 3 groups simultaneously, they found that the most accurate method was SVM, which attained 89.5% sensitivity for the detection of subclinical keratoconus, and the results were validated using 10-fold cross-validation. The limitations of this study include the use of the CASIA ectasia screening index (ESI) for the classification of the severity of keratoconus, which may not agree with clinical diagnosis, and that the analyzed parameters are closely tied to the CASIA device, which limits generalizability to other systems.

Unsupervised Learning

Unsupervised learning represents a distinct approach to the detection of subclinical keratoconus by attempting to identify groups of similar eyes without prelabeled data. Yousefi et al [61] used a 2-step approach that combined dimensionality reduction and density-based clustering to cluster a cohort of 3156 eyes categorized according to the ESI index as either normal, keratoconus, and subclinical (FFKC) keratoconus. They included 420 topography, elevation, and pachymetry parameters, and the algorithm produced 4 clusters of eyes with similar characteristics. When comparing their results with a reference standard (ESI), the model did not create a distinct grouping that separated the subclinical eyes from other eyes (sensitivity 88% and specificity 14%), suggesting poor correlation when compared with ESI alone. Furthermore, they did not compare their results with clinically labeled data.

Research Question 3: How Was Algorithm Validation Handled Among the Selected Manuscripts?

Although most studies performed internal validation by splitting the original data set into training and test sets, we identified 5 replication papers that validated a published model on a new

data set [48,51,52,56,63]. Ruiz Hidalgo et al [56] verified their SVM technique presented in 2016 [55]. The authors found that when using the Antwerp University Hospital classification, there was approximately 18% decrease in sensitivity, whereas when using the Rothschild classification, there was approximately 21% increase in sensitivity. These discrepancies highlight the problems associated with subjective classification and the absence of ground truth. Furthermore, when multiple groups were included in the analysis; that is, normal, keratoconus, subclinical keratoconus, and postrefractive surgery eyes, it was noted that the accuracy decreased from 93.1% in discriminating normal from FFKC to 88.8%. However, this paper presented the most comprehensive methodology because the authors not only verified their results on a new sample population with multiple target classes but also compared their results with other methods and included 2 subjective reference standards.

Buhren et al [51] validated their model defined in 2010 [84]. When comparing their discriminant function derived from anterior and posterior corneal surface wavefront data, they reported approximately 22% decrease in sensitivity and approximately 9% decrease in specificity. This decrease was likely caused by overfitting in the original study. Saad et al [48] and Chan et al [52] validated the same discriminant analysis model presented by Saad et al [77]. Saad et al [48] reported sensitivity (92%) and specificity (96%), roughly in line with their previous study, which indicates that their method is reliable and does not suffer from overfitting. Chan et al [52] validated the original model in patients from different ethnic backgrounds (Asian). They reported approximately 21% decrease in sensitivity, which they attributed to overfitting in the original study; however, their specificity was almost equivalent. Steinberg et al [63] validated the work presented by Ambrosio et al [58]. They reported approximately 27% decrease in sensitivity and approximately 13% decrease in specificity when applying the same thresholds.

Bias Assessment

In general, patient selection was found to have a high risk of bias (19/26, 73% of studies) because most studies were case-control (thus susceptible to selection bias) and did not use consecutive or random samples. Multimedia Appendix 5 [14,47-71] contains the results of applying the QUADAS-2 tool when considering the risk of bias. The index test was also generally found to have a high risk of bias (21/26, 81% of studies) because of the lack of external validation. As there is no gold standard for subclinical keratoconus diagnosis, we could not assess the bias for the reference standard; therefore, all papers were marked as unclear. Finally, patient flow was found to have a low risk of bias (21/26, 81% of studies) because although chronological information was sparse, the same analysis was usually applied to all patients.

Discussion

Research Question 1: What Input Data Types Have Been Used Within Subclinical Keratoconus Detection Algorithms and How Have They Performed?

The data most commonly used for building subclinical keratoconus detection algorithms are numeric keratometry or pachymetry parameters; hence, according to our review, algorithms based on these tend to have the highest performance. These parameters are derived from a variety of imaging systems and devices and are then incorporated into different combinations to build a classification system or an index. Inevitably, individual systems produce parameters that may not be comparable across devices, and for proprietary reasons, the raw data are generally not available to derive these parameters. Therefore, comparison or replication across systems is difficult. Heat maps provide a visual representation of either corneal elevation, pachymetry, or curvature, which are helpful for the visual interpretation of results. However, heat maps require interpolation or extrapolation of data, which may introduce inaccuracies when included in the model. To the best of our knowledge, there are no studies that have analyzed actual pixel-level corneal imaging data (Scheimpflug or AS-OCT), probably because access to these data is restricted to commercial machines such as the Pentacam, which impedes bulk export to train machine learning algorithms.

We also noted that many studies do not incorporate details of patient demographics and associated diseases, such as age, sex, ethnicity, and atopy, which can influence the risk of developing keratoconus. Incorporating these data into these models may help define the population to which an algorithm applies, particularly as there are phenotypic indices that an algorithm can identify from images that humans cannot identify by manual inspection [85].

Research Question 2: What Machine Learning Algorithms Have Been Used for Subclinical Keratoconus Detection and How Have They Performed?

Subclinical keratoconus studies typically involve univariate or multivariate analyses. For univariate studies, receiver operating characteristic analysis is performed, as each parameter is included to quantify their diagnostic ability. However, because none of the univariate studies we identified performed an out-of-sample validation, they were all excluded. For multivariate studies, machine learning is used to create a detection model using multiple parameters. These algorithms have already demonstrated comparable performance to experienced ophthalmologists in the identification of retinopathy of prematurity [86] and retinal disease progression [87]. Machine learning–based research into the detection of subclinical keratoconus has largely focused on supervised learning techniques, such as decision trees, SVM, logistic regression, discriminant analysis, NNs, and CNNs. Logistic regression may be superior to NNs when parameters from a single imaging modality are considered [14,68], with a potentially greater role for NNs when a large number of potentially interacting parameters are combined, such as for multiple imaging

modalities [68]. Unsupervised learning has also been evaluated for the detection of subclinical keratoconus, although it relies on identifying patterns in large amounts of data; hence, it may not translate to a different data set of a different size and with different properties. In addition, with the exception of the study by Yousefi [61], none of the papers provided access to the source code for their algorithms or a description of the hyperparameters, which makes it difficult to reproduce and validate the results with external data sets.

Research Question 3: How Was Algorithm Validation Handled Among the Selected?

We excluded papers that did not include a validation arm for the study, and the vast majority of initially identified studies did not appropriately validate their results. For any type of automatic classifier, validating the results on a data set distinct from the trained set is critical in determining the generalizability of the model to other data sets. With the exception of the studies by Saad et al [48] and Hidalgo et al [56], it is clear that studies attempting to validate a prior method reported significant decreases in sensitivity and specificity in comparison with their original results. This shows that even when techniques such as cross-validation are performed, the best method for validation is an independent out-of-sample data set, and its absence may introduce bias. Ideally, this external data set would be larger and more representative of the general population.

Strengths and Limitations

The primary strength of this study is that we present a comprehensive review of all studies published in English between January 1, 2010, and October 31, 2020, on the use of machine learning for the detection of subclinical keratoconus. Our focus on the detection of subclinical keratoconus addresses an important unmet clinical need for an effective machine-based technique to identify keratoconus at its earliest stage. This would move us closer to potential screening without significant demands on clinicians and clinical services. Subclinical disease diagnosis is more challenging than the detection of advanced disease, where the opportunity to prevent progression has already been lost. In this respect, our review builds on recent clinical trials of CXL to prevent keratoconus progression in children and young adults [15,88,89]. To present a balanced and comprehensive overview, we have combined the expertise of computer scientists (HM and NP) familiar with the development of machine learning for clinical medicine with the input from clinicians (JPOL, DG, and ST) who are experienced in keratoconus management. We have considered and compared the literature in terms of both clinical input data and machine learning methodology, which allows the reader to gain a wider perspective of the problem.

However, there are limitations to our search methods and inclusion criteria. As with any systematic review, articles that did not include the relevant key terms or were not appropriately indexed by the literature databases may have been missed. When considering our inclusion criteria based on subclinical disease, some studies may have been missed because of a lack of consensus on definition. In addition, where there was no form of validation, we excluded the study; thus, our results represent only the articles that have some degree of generalizability.

A further limitation is the difficulty in comparing the performance of the approaches described in the manuscripts; direct comparisons were not possible because of the variation of multiple study design factors such as subclinical disease definition, parameter choice, data set source, and machine learning algorithm. Finally, a limitation regarding case definition that applies to all studies is the uncertainty in the relationship between subclinical keratoconus and other nonprogressive abnormalities of corneal shape.

Challenges and Future Directions

Our systematic review identified several challenges from the literature and avenues for future research.

Case Definition, Gold Standard, and Ground Truth

Precise comparisons between the results of publications are problematic because of the ambiguous definition of early keratoconus and the absence of a gold standard examination technique. The most common definition of subclinical keratoconus is an eye with topographic findings that is at least suspicious of keratoconus and with confirmed keratoconus in the fellow eye. FFKC is usually defined as an eye that has both normal topography and slit-lamp examination but with keratoconus in the fellow eye [44]. With this differentiation, subclinical keratoconus will be easier to detect than FFKC, and studies using the former definition are likely to produce more accurate results because the problem becomes easier to solve. The problems of making statistical comparisons in the absence of a gold standard have been discussed extensively by Umemneku et al [90]. The authors suggest that latent class analysis, composite reference standards, or expert panel analysis may be appropriate in these circumstances.

Even if a precise definition of early subclinical keratoconus was established, the absence of ground truth data is relevant when evaluating the precision of data acquisition. For example, measurements of keratoconus taken by different operators or repeated on different days may lead to variations in the results. Flynn et al [91] found that keratometric measurements from Scheimpflug images (Pentacam) were more reproducible in early keratoconus (mean central K ≤53 D) compared with those in more advanced keratoconus (mean central K>53 D), although a cohort with subclinical keratoconus was not included. In contrast, Yang et al [92] found that biomechanical parameters (Corvis ST) had acceptable repeatability in both normal and keratoconus eyes.

Another issue we identified when comparing studies was the variation in the number of groups that were classified. The studies often started with multiple groups (usually 3, eg, FFKC, keratoconus, and normal); however, 21 papers chose to report their accuracy results from a model trained to classify between just 2 groups (eg, FFKC and normal), whereas 5 papers reported results for classifying between all groups. Classifying all groups is a more realistic clinical scenario, but it presents a more challenging problem because the features of the different groups can overlap. Complete details of the number of groups associated with the accuracy results are presented in Multimedia Appendix 2.

Study Size and Statistical Power

The size of the study is critical when developing a reliable detection system. In particular, the accuracy of machine learning models is directly related to the amount of training data. When considering eyes with subclinical keratoconus, only 2 studies included more than 500 eyes [61,71]. None of the papers included a priori power calculations to estimate the size of the cohort to be studied.

Study Design

None of the reported studies evaluated the performance of their method against masked observers; thus, they may introduce detection bias. The initial classification is often made by considering the fellow eye with keratoconus as a factor in the decision-making process, whereas the algorithm does not have this information. Hence, it would be interesting to design a study where, having already decided on the ground truth diagnosis, a new clinician is asked to evaluate the eye using the same information as the algorithm (ie, only the images or parameters). This situation is closest to real-life screening where a prospective patient (without a history of keratoconus in either eye) is examined for risk of keratoconus.

Subclinical keratoconus is, by definition, the least affected eye of highly asymmetrical keratoconus. An assumption is that any parameters of subclinical disease that differ from the values for normal corneas are the result of keratoconus. However, it has not been demonstrated prospectively that all eyes in such a cohort will progress to the clinical disease state. Although true unilateral keratoconus is thought not to exist [37], this has not been proven, and it is possible that some eyes with subclinical keratoconus are not at risk of progression and that some of the abnormal parameters in this group are not the result of keratoconus. It would be valuable to conduct a prospective study in which eyes that do not develop clinical keratoconus over time are used as lower-risk examples.

External Validation and Generalizability to Real-world Data

To be useful, it is essential that a detection algorithm can generalize beyond the limited data set from which the model was developed and benchmarked, which requires external validation in out-of-sample data sets. The creation of a large open-source data set of keratoconus images could serve as a reference standard to develop a benchmark for external validation. We also recommend that journals adhere to the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis guidelines so that all published methods are externally validated. When generalizing to external data sets the source and quality of the data should be considered. Data from a referral hospital may not represent the general population, who might be the target for screening programs, with an underrepresentation of eyes with mild disease.

Other Challenges

There are several other considerations, such as keratoconus progression and the translation of a detection algorithm into a medical device that can be implemented in the real world, but these issues are beyond the scope of this review. Nevertheless, these points are discussed in Multimedia Appendix 6 [37,39,93-102].

New Avenues of Research

On the basis of the results of this review, there is a need for further fundamental research, particularly for analysis based on the raw pixel values of corneal imaging rather than only derived parameters. Furthermore, a multimodal solution could be developed by combining these raw images with other parameters, such as biomechanical, demographic, and genetic data. Demographic data such as age, sex, ethnicity, and allergic eye disease are known risk factors for progressive keratoconus, and a family history of keratoconus is also a risk factor that should also be included in diagnostic algorithms. Environmental risk factors, including eye rubbing, have been associated with keratoconus progression, although eye rubbing is difficult to quantify. A genetic predisposition to keratoconus is supported by heritability studies in twins, linkage analysis in families, and population-level genome-wide association studies [103]. From these studies, genetic risk scores have been derived, which could be included in machine learning models for the detection of subclinical keratoconus. Ideally, a prospective study should be performed in a large cohort of young (<30 years of age) patients with subclinical keratoconus to monitor disease progression. Training should be conducted on large data sets with the explicit aim of detecting subclinical keratoconus, and the resulting model should be externally validated on a new data set. Finally, a range of machine learning techniques should be applied to the same data set along with detailed comparison statistics.

Conclusions

We have conducted the most comprehensive review to date on machine learning algorithms for the detection of subclinical keratoconus. Early detection of keratoconus to enable treatment and prevent sight loss is a public health priority, and the use of machine learning algorithms has the potential to make the diagnostic process more efficient and widely available. We have summarized the relevant publications in terms of their input data and the choice of algorithm and identified whether studies performed appropriate validation. We have identified the challenges of obtaining accurate data sets for training machine learning algorithms and the need for a consistent, objective, and agreed definition of subclinical keratoconus. New avenues of research have been identified that combine multimodal source data with biomechanical, demographic, and genomic data. Defining disease progression and modeling progression to the point where there is sight loss are areas that may benefit from further research. We believe this up-to-date review is important to enable researchers, clinicians, and public health policymakers to understand the current state of the research and provide guidance for future health service planning.

Acknowledgments

HM is funded by Moorfields Eye Charity (GR001147). NP is funded by Moorfields Eye Charity Career Development Award (R190031A). Moorfields Eye Charity is supported in part by the National Institute for Health Research Biomedical Research Centre based at Moorfields Eye Hospital National Health Service Foundation Trust and University College London Institute of Ophthalmology. PL is supported by Progress Q26/LF1 and UNCE 204064. ST and DG acknowledge that a proportion of their financial support is from the Department of Health through the award made by the National Institute for Health Research to Moorfields Eye Hospital National Health Service Foundation Trust and University College London Institute of Ophthalmology for a Specialist Biomedical Research Centre for Ophthalmology.

The views expressed are those of the authors and not necessarily those of the National Health Service, the National Institute for Health Research, or the UK Department of Health and Social Care.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Grading systems and indices.
[DOC File , 55 KB - medinform_v9i12e27363_app1.doc]

Multimedia Appendix 2
Supplementary details for the 26 published studies that included the use of machine learning for the detection of subclinical keratoconus.
[XLS File (Microsoft Excel File), 54 KB - medinform_v9i12e27363_app2.xls]

Multimedia Appendix 3
Sensitivity and specificity were plotted for the 26 published studies that included the use of machine learning for the detection of subclinical keratoconus. In the left-hand chart, the results were grouped by diagnosis criteria. A: clinically normal, topographically abnormal. B: Fellow eye of diagnosed keratoconus, clinically normal, topographically normal. C: Fellow eye of diagnosed keratoconus, clinically normal, topographically abnormal. Although there is no obvious pattern relating to diagnostic criteria, the largest outliers belong to group A, suggesting that using a fellow eye with keratoconus may lead to a better detection system. In the right-hand chart, the results are grouped according to the imaging system. No obvious pattern can be seen in the results, suggesting that the choice of imaging system is unrelated to the detection system accuracy.
[PNG File , 56 KB - medinform_v9i12e27363_app3.png]

Multimedia Appendix 4
Heat maps of an advanced keratoconus eye derived from Scheimpflug corneal imaging using the Pentacam device.
[PNG File , 427 KB - medinform_v9i12e27363_app4.png]

Multimedia Appendix 5
Results of applying the QUADAS (Quality Assessment of Diagnostic Accuracy Studies)-2 bias assessment tool including responses to tailored signaling questions.
[XLS File (Microsoft Excel File), 33 KB - medinform_v9i12e27363_app5.xls]

Multimedia Appendix 6
Other challenges, such as keratoconus progression and translational considerations, have been identified for the detection of subclinical keratoconus using machine learning.
[DOCX File , 97 KB - medinform_v9i12e27363_app6.docx]

References

1. Mas Tur V, MacGregor C, Jayaswal R, O'Brart D, Maycock N. A review of keratoconus: diagnosis, pathophysiology, and genetics. Surv Ophthalmol 2017;62(6):770-783. [doi: 10.1016/j.survophthal.2017.06.009] [Medline: 28688894]
2. Davidson AE, Hayes S, Hardcastle AJ, Tuft SJ. The pathogenesis of keratoconus. Eye (Lond) 2014 Feb;28(2):189-195 [FREE Full text] [doi: 10.1038/eye.2013.278] [Medline: 24357835]
3. Godefrooij DA, de Wit GA, Uiterwaal CS, Imhof SM, Wisse RP. Age-specific incidence and prevalence of keratoconus: a nationwide registration study. Am J Ophthalmol 2017 Mar;175:169-172. [doi: 10.1016/j.ajo.2016.12.015] [Medline: 28039037]

4. Chan E, Chong EW, Lingham G, Stevenson LJ, Sanfilippo PG, Hewitt AW, et al. Prevalence of keratoconus based on Scheimpflug imaging: the Raine study. Ophthalmology 2021 Apr;128(4):515-521. [doi: 10.1016/j.ophtha.2020.08.020] [Medline: 32860813]
5. Papali i-Curtin AT, Cox R, Ma T, Woods L, Covello A, Hall RC. Keratoconus prevalence among high school students in New Zealand. Cornea 2019 Nov;38(11):1382-1389. [doi: 10.1097/ICO.0000000000002054] [Medline: 31335534]
6. Ferdi AC, Nguyen V, Gore DM, Allan BD, Rozema JJ, Watson SL. Keratoconus natural progression: a systematic review and meta-analysis of 11 529 eyes. Ophthalmology 2019 Jul;126(7):935-945. [doi: 10.1016/j.ophtha.2019.02.029] [Medline: 30858022]
7. Tuft SJ, Moodaley LC, Gregory WM, Davison CR, Buckley RJ. Prognostic factors for the progression of keratoconus. Ophthalmology 1994 Mar;101(3):439-447. [doi: 10.1016/s0161-6420(94)31313-3] [Medline: 8127564]
8. Pearson AR, Soneji B, Sarvananthan N, Sandford-Smith JH. Does ethnic origin influence the incidence or severity of keratoconus? Eye (Lond) 2000 Aug;14 (Pt 4):625-628. [doi: 10.1038/eye.2000.154] [Medline: 11040911]
9. Downie LE, Lindsay RG. Contact lens management of keratoconus. Clin Exp Optom 2015 Jul;98(4):299-311 [FREE Full text] [doi: 10.1111/cxo.12300] [Medline: 26104589]
10. Steinberg J, Bußmann N, Frings A, Katz T, Druchkiv V, Linke SJ. Quality of life in stable and progressive 'early-stage' keratoconus patients. Acta Ophthalmol 2021 Mar;99(2):e196-e201. [doi: 10.1111/aos.14564] [Medline: 32914586]
11. Saunier V, Mercier A, Gaboriau T, Malet F, Colin J, Fournié P, et al. Vision-related quality of life and dependency in French keratoconus patients: impact study. J Cataract Refract Surg 2017 Dec;43(12):1582-1590. [doi: 10.1016/j.jcrs.2017.08.024] [Medline: 29335104]
12. Gore DM, Watson MP, Tuft SJ. Permanent visual loss in eyes with keratoconus. Acta Ophthalmol 2014 May;92(3):e244-e245 [FREE Full text] [doi: 10.1111/aos.12253] [Medline: 23910953]
13. Kelly T, Williams KA, Coster DJ, Australian Corneal Graft Registry. Corneal transplantation for keratoconus: a registry study. Arch Ophthalmol 2011 Jun;129(6):691-697. [doi: 10.1001/archophthalmol.2011.7] [Medline: 21320951]
14. Cao K, Verspoor K, Sahebjada S, Baird PN. Evaluating the performance of various machine learning algorithms to detect subclinical keratoconus. Trans Vis Sci Tech 2020 Apr 24;9(2):24. [doi: 10.1167/tvst.9.2.24] [Medline: 32818085]
15. Wittig-Silva C, Chan E, Islam FM, Wu T, Whiting M, Snibson GR. A randomized, controlled trial of corneal collagen cross-linking in progressive keratoconus: three-year results. Ophthalmology 2014 Apr;121(4):812-821. [doi: 10.1016/j.ophtha.2013.10.028] [Medline: 24393351]
16. Caporossi A, Mazzotta C, Baiocchi S, Caporossi T. Long-term results of riboflavin ultraviolet a corneal collagen cross-linking for keratoconus in Italy: the Siena eye cross study. Am J Ophthalmol 2010 Apr;149(4):585-593. [doi: 10.1016/j.ajo.2009.10.021] [Medline: 20138607]
17. O'Brart DP, Chan E, Samaras K, Patel P, Shah SP. A randomised, prospective study to investigate the efficacy of riboflavin/ultraviolet A (370 nm) corneal collagen cross-linkage to halt the progression of keratoconus. Br J Ophthalmol 2011 Nov;95(11):1519-1524. [doi: 10.1136/bjo.2010.196493] [Medline: 21349938]
18. Gore DM, Shortt AJ, Allan BD. New clinical pathways for keratoconus. Eye (Lond) 2013 Mar;27(3):329-339 [FREE Full text] [doi: 10.1038/eye.2012.257] [Medline: 23258309]
19. Koller T, Mrochen M, Seiler T. Complication and failure rates after corneal crosslinking. J Cataract Refract Surg 2009 Aug;35(8):1358-1362. [doi: 10.1016/j.jcrs.2009.03.035] [Medline: 19631120]
20. Gore DM, Leucci MT, Koay S, Kopsachilis N, Nicolae MN, Malandrakis MI, et al. Accelerated pulsed high-fluence corneal cross-linking for progressive keratoconus. Am J Ophthalmol 2021 Jan;221:9-16. [doi: 10.1016/j.ajo.2020.08.021] [Medline: 32818448]
21. Salmon HA, Chalk D, Stein K, Frost NA. Cost effectiveness of collagen crosslinking for progressive keratoconus in the UK NHS. Eye (Lond) 2015 Nov;29(11):1504-1511 [FREE Full text] [doi: 10.1038/eye.2015.151] [Medline: 26315704]
22. Lindstrom RL, Berdahl JP, Donnenfeld ED, Thompson V, Kratochvil D, Wong C, et al. Corneal cross-linking versus conventional management for keratoconus: a lifetime economic model. J Med Econ 2021;24(1):410-420. [doi: 10.1080/13696998.2020.1851556] [Medline: 33210975]
23. Godefrooij DA, Mangen MJ, Chan E, O'Brart DP, Imhof SM, de Wit GA, et al. Cost-effectiveness analysis of corneal collagen crosslinking for progressive keratoconus. Ophthalmology 2017 Oct;124(10):1485-1495. [doi: 10.1016/j.ophtha.2017.04.011] [Medline: 28532974]
24. Kreps EO, Claerhout I, Koppen C. Diagnostic patterns in keratoconus. Cont Lens Anterior Eye 2021 Jun;44(3):101333. [doi: 10.1016/j.clae.2020.05.002] [Medline: 32448765]
25. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine learning has arrived!. Ophthalmology 2017 Dec;124(12):1726-1728. [doi: 10.1016/j.ophtha.2017.08.046] [Medline: 29157423]
26. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. USA: Springer; 2009.
27. Russell S, Norvig P. Artificial Intelligence: A Modern Approach. USA: Prentice Hall; 2020.
28. Tong Y, Lu W, Yu Y, Shen Y. Application of machine learning in ophthalmic imaging modalities. Eye Vis (Lond) 2020;7:22 [FREE Full text] [doi: 10.1186/s40662-020-00183-6] [Medline: 32322599]

29. Lin SR, Ladas JG, Bahadur GG, Al-Hashimi S, Pineda R. A review of machine learning techniques for keratoconus detection and refractive surgery screening. Semin Ophthalmol 2019;34(4):317-326. [doi: 10.1080/08820538.2019.1620812] [Medline: 31304857]
30. Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics). USA: Springer-Verlag; 2007.
31. Goodfellow I, Bengio Y, Courville A. Deep Learning (Adaptive Computation and Machine Learning Series). USA: MIT Press; 2017:800.
32. Ting DS, Pasquale LR, Peng L, Campbell JP, Lee AY, Raman R, et al. Artificial intelligence and deep learning in ophthalmology. Br J Ophthalmol 2019 Feb;103(2):167-175 [FREE Full text] [doi: 10.1136/bjophthalmol-2018-313173] [Medline: 30361278]
33. Korot E, Guan Z, Ferraz D, Wagner SK, Zhang G, Liu X, et al. Code-free deep learning for multi-modality medical image classification. Nat Mach Intell 2021 Mar 01;3(4):288-298. [doi: 10.1038/s42256-021-00305-2]
34. Wang Y, Zhao Y, Therneau TM, Atkinson EJ, Tafti AP, Zhang N, et al. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. J Biomed Inform 2020 Feb;102:103364 [FREE Full text] [doi: 10.1016/j.jbi.2019.103364] [Medline: 31891765]
35. Pontikos N. Normalisation and Clustering Methods Applied to Association Studies in Type 1 Diabetes. UK: Cambridge University; 2015.
36. -. Grading diabetic retinopathy from stereoscopic color fundus photographs--an extension of the modified Airlie House classification. ETDRS report number 10. Early Treatment Diabetic Retinopathy Study Research Group. Ophthalmology 1991 May;98(5 Suppl):786-806. [Medline: 2062513]
37. Gomes JA, Tan D, Rapuano CJ, Belin MW, Ambrósio R, Guell JL, Group of Panelists for the Global Delphi Panel of KeratoconusEctatic Diseases. Global consensus on keratoconus and ectatic diseases. Cornea 2015 Apr;34(4):359-369. [doi: 10.1097/ICO.0000000000000408] [Medline: 25738235]
38. Amsler M. Kératocône classique et kératocône fruste; arguments unitaires. Ophthalmologica 1946;111(2-3):96-101. [doi: 10.1159/000300309] [Medline: 20275788]
39. Belin MW, Duncan JK. Keratoconus: the ABCD grading system. Klin Monbl Augenheilkd 2016 Jun;233(6):701-707. [doi: 10.1055/s-0042-100626] [Medline: 26789119]
40. Independent population validation of the belinambrosio enhanced ectasia display implications for keratoconus studies and screening. Enhanced Screening for Ectasia Detection based on Scheimpflug Tomography and Biomachanical evaluations. 2014. URL: https://tinyurl.com/5yywh3fm [accessed 2021-11-22]
41. Hashemi H, Beiranvand A, Yekta A, Maleki A, Yazdani N, Khabazkhoob M. Pentacam top indices for diagnosing subclinical and definite keratoconus. J Curr Ophthalmol 2016 Mar;28(1):21-26 [FREE Full text] [doi: 10.1016/j.joco.2016.01.009] [Medline: 27239598]
42. Pentacam®. OCULUS. URL: https://www.pentacam.com/int/opticianoptometrist-without-pentacamr/models/pentacamr/core-functions.html [accessed 2021-11-22]
43. Ocular Response Analyzer® G3. Reichert Technologies. URL: https://www.reichert.com/products/ocular-response-analyzer-g3 [accessed 2021-11-22]
44. Henriquez MA, Hadid M, Izquierdo L. A systematic review of subclinical keratoconus and forme fruste keratoconus. J Refract Surg 2020 Apr 01;36(4):270-279. [doi: 10.3928/1081597X-20200212-03] [Medline: 32267959]
45. 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 [FREE Full text] [doi: 10.1371/journal.pmed.1000097] [Medline: 19621072]
46. Krumeich JH, Daniel J, Knülle A. Live-epikeratophakia for keratoconus. J Cataract Refract Surg 1998 Apr;24(4):456-463. [doi: 10.1016/s0886-3350(98)80284-8] [Medline: 9584238]
47. Arbelaez MC, Versaci F, Vestri G, Barboni P, Savini G. Use of a support vector machine for keratoconus and subclinical keratoconus detection by topographic and tomographic data. Ophthalmology 2012 Nov;119(11):2231-2238. [doi: 10.1016/j.ophtha.2012.06.005] [Medline: 22892148]
48. Saad A, Gatinel D. Validation of a new scoring system for the detection of early forme of keratoconus. Int J Keratoconus Ectatic Corneal Dis 2012 May;1(2):100-108. [doi: 10.5005/JP-JOURNALS-10025-1019]
49. Smadja D, Touboul D, Cohen A, Doveh E, Santhiago MR, Mello GR, et al. Detection of subclinical keratoconus using an automated decision tree classification. Am J Ophthalmol 2013 Aug;156(2):237-46.e1. [doi: 10.1016/j.ajo.2013.03.034] [Medline: 23746611]
50. Ramos-López D, Martínez-Finkelshtein A, Castro-Luna GM, Burguera-Gimenez N, Vega-Estrada A, Piñero D, et al. Screening subclinical keratoconus with placido-based corneal indices. Optom Vis Sci 2013 Apr;90(4):335-343. [doi: 10.1097/OPX.0b013e3182843f2a] [Medline: 23376898]
51. Bühren J, Schäffeler T, Kohnen T. Validation of metrics for the detection of subclinical keratoconus in a new patient collective. J Cataract Refract Surg 2014 Feb;40(2):259-268. [doi: 10.1016/j.jcrs.2013.07.044] [Medline: 24360499]
52. Chan C, Ang M, Saad A, Chua D, Mejia M, Lim L, et al. Validation of an objective scoring system for forme fruste keratoconus detection and post-lasik ectasia risk assessment in asian eyes. Cornea 2015 Sep;34(9):996-1004. [doi: 10.1097/ICO.0000000000000529] [Medline: 26165793]

53. Kovács I, Miháltz K, Kránitz K, Juhász É, Takács Á, Dienes L, et al. Accuracy of machine learning classifiers using bilateral data from a Scheimpflug camera for identifying eyes with preclinical signs of keratoconus. J Cataract Refract Surg 2016 Feb;42(2):275-283. [doi: 10.1016/j.jcrs.2015.09.020] [Medline: 27026453]
54. Saad A, Gatinel D. Combining placido and corneal wavefront data for the detection of forme fruste keratoconus. J Refract Surg 2016 Aug 01;32(8):510-516. [doi: 10.3928/1081597X-20160523-01] [Medline: 27505311]
55. Ruiz Hidalgo I, Rodriguez P, Rozema JJ, Ní Dhubhghaill S, Zakaria N, Tassignon M, et al. Evaluation of a machine-learning classifier for keratoconus detection based on scheimpflug tomography. Cornea 2016 Jun;35(6):827-832. [doi: 10.1097/ICO.0000000000000834] [Medline: 27055215]
56. Ruiz Hidalgo I, Rozema JJ, Saad A, Gatinel D, Rodriguez P, Zakaria N, et al. Validation of an objective keratoconus detection system implemented in a scheimpflug tomographer and comparison with other methods. Cornea 2017 Jun;36(6):689-695. [doi: 10.1097/ICO.0000000000001194] [Medline: 28368992]
57. Xu Z, Li W, Jiang J, Zhuang X, Chen W, Peng M, et al. Characteristic of entire corneal topography and tomography for the detection of sub-clinical keratoconus with Zernike polynomials using Pentacam. Sci Rep 2017 Nov 28;7(1):16486 [FREE Full text] [doi: 10.1038/s41598-017-16568-y] [Medline: 29184086]
58. Ambrósio R, Lopes BT, Faria-Correia F, Salomão MQ, Bühren J, Roberts CJ, et al. Integration of scheimpflug-based corneal tomography and biomechanical assessments for enhancing ectasia detection. J Refract Surg 2017 Jul 01;33(7):434-443 [FREE Full text] [doi: 10.3928/1081597X-20170426-02] [Medline: 28681902]
59. Sideroudi H, Labiris G, Georgantzoglou K, Ntonti P, Siganos C, Kozobolis V. Fourier analysis algorithm for the posterior corneal keratometric data: clinical usefulness in keratoconus. Ophthalmic Physiol Opt 2017 Jul;37(4):460-466. [doi: 10.1111/opo.12386] [Medline: 28656673]
60. Francis M, Pahuja N, Shroff R, Gowda R, Matalia H, Shetty R, et al. Waveform analysis of deformation amplitude and deflection amplitude in normal, suspect, and keratoconic eyes. J Cataract Refract Surg 2017 Oct;43(10):1271-1280. [doi: 10.1016/j.jcrs.2017.10.012] [Medline: 29120713]
61. Yousefi S, Yousefi E, Takahashi H, Hayashi T, Tampo H, Inoda S, et al. Keratoconus severity identification using unsupervised machine learning. PLoS One 2018 Nov 6;13(11):e0205998 [FREE Full text] [doi: 10.1371/journal.pone.0205998] [Medline: 30399144]
62. Lopes BT, Ramos IC, Salomão MQ, Guerra FP, Schallhorn SC, Schallhorn JM, et al. Enhanced tomographic assessment to detect corneal ectasia based on artificial intelligence. Am J Ophthalmol 2018 Nov;195:223-232. [doi: 10.1016/j.ajo.2018.08.005] [Medline: 30098348]
63. Steinberg J, Siebert M, Katz T, Frings A, Mehlan J, Druchkiv V, et al. Tomographic and biomechanical scheimpflug imaging for keratoconus characterization: a validation of current indices. J Refract Surg 2018 Dec 01;34(12):840-847. [doi: 10.3928/1081597X-20181012-01] [Medline: 30540367]
64. Issarti I, Consejo A, Jiménez-García M, Hershko S, Koppen C, Rozema JJ. Computer aided diagnosis for suspect keratoconus detection. Comput Biol Med 2019 Jun;109:33-42. [doi: 10.1016/j.compbiomed.2019.04.024] [Medline: 31035069]
65. Chandapura R, Salomão MQ, Ambrósio R, Swarup R, Shetty R, Sinha Roy A. Bowman's topography for improved detection of early ectasia. J Biophotonics 2019 Oct;12(10):e201900126. [doi: 10.1002/jbio.201900126] [Medline: 31152630]
66. Xie Y, Zhao L, Yang X, Wu X, Yang Y, Huang X, et al. Screening candidates for refractive surgery with corneal tomographic-based deep learning. JAMA Ophthalmol 2020 May 01;138(5):519-526 [FREE Full text] [doi: 10.1001/jamaophthalmol.2020.0507] [Medline: 32215587]
67. Kuo B, Chang W, Liao T, Liu F, Liu H, Chu H, et al. Keratoconus screening based on deep learning approach of corneal topography. Transl Vis Sci Technol 2020 Sep 25;9(2):53 [FREE Full text] [doi: 10.1167/tvst.9.2.53] [Medline: 33062398]
68. Shi C, Wang M, Zhu T, Zhang Y, Ye Y, Jiang J, et al. Machine learning helps improve diagnostic ability of subclinical keratoconus using Scheimpflug and OCT imaging modalities. Eye Vis (Lond) 2020 Sep 10;7:48 [FREE Full text] [doi: 10.1186/s40662-020-00213-3] [Medline: 32974414]
69. Toprak I, Cavas F, Velázquez JS, Alio Del Barrio JL, Alio JL. Subclinical keratoconus detection with three-dimensional (3-D) morphogeometric and volumetric analysis. Acta Ophthalmol 2020 Dec;98(8):e933-e942. [doi: 10.1111/aos.14433] [Medline: 32410342]
70. Issarti I, Consejo A, Jiménez-García M, Kreps EO, Koppen C, Rozema JJ. Logistic index for keratoconus detection and severity scoring (Logik). Comput Biol Med 2020 Jul;122:103809. [doi: 10.1016/j.compbiomed.2020.103809] [Medline: 32658727]
71. Lavric A, Popa V, Takahashi H, Yousefi S. Detecting keratoconus from corneal imaging data using machine learning. IEEE Access 2020 Aug 12;8:149113-149121. [doi: 10.1109/access.2020.3016060]
72. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. Indian J Ophthalmol 2008;56(1):45-50 [FREE Full text] [doi: 10.4103/0301-4738.37595] [Medline: 18158403]
73. Whiting PF, Rutjes AW, 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 [FREE Full text] [doi: 10.7326/0003-4819-155-8-201110180-00009] [Medline: 22007046]
74. Lawless MA, Hodge C. Wavefront's role in corneal refractive surgery. Clin Exp Ophthalmol 2005 Apr;33(2):199-209. [doi: 10.1111/j.1442-9071.2005.00994.x] [Medline: 15807834]

75. Toprak I, Vega A, Alió Del Barrio JL, Espla E, Cavas F, Alió JL. Diagnostic value of corneal epithelial and stromal thickness distribution profiles in forme fruste keratoconus and subclinical keratoconus. Cornea 2021 Jan;40(1):61-72. [doi: 10.1097/ICO.0000000000002435] [Medline: 32769675]
76. Belin MW, Khachikian SS. Keratoconus / ectasia detection with the oculus pentacam: belin / ambrósio enhanced ectasia display. Highlights Ophthalmol 2008;35(6):5-12. [doi: 10.5005/jp/books/11830_8]
77. Saad A, Gatinel D. Topographic and tomographic properties of forme fruste keratoconus corneas. Invest Ophthalmol Vis Sci 2010 Nov;51(11):5546-5555. [doi: 10.1167/iovs.10-5369] [Medline: 20554609]
78. Rozema JJ, Rodriguez P, Ruiz Hidalgo I, Navarro R, Tassignon M, Koppen C. SyntEyes KTC: higher order statistical eye model for developing keratoconus. Ophthalmic Physiol Opt 2017 May;37(3):358-365. [doi: 10.1111/opo.12369] [Medline: 28303580]
79. Hashemi H, Beiranvand A, Yekta A, Asharlous A, Khabazkhoob M. Biomechanical properties of early keratoconus: suppressed deformation signal wave. Cont Lens Anterior Eye 2017 Apr;40(2):104-108. [doi: 10.1016/j.clae.2016.12.004] [Medline: 27956045]
80. Galletti JD, Ruiseñor Vázquez PR, Fuentes Bonthoux F, Pförtner T, Galletti JG. Multivariate analysis of the ocular response analyzer's corneal deformation response curve for early keratoconus detection. J Ophthalmol 2015;2015:496382 [FREE Full text] [doi: 10.1155/2015/496382] [Medline: 26075085]
81. Barsam A, Petrushkin H, Brennan N, Bunce C, Xing W, Foot B, et al. Acute corneal hydrops in keratoconus: a national prospective study of incidence and management. Eye (Lond) 2015 Apr;29(4):469-474 [FREE Full text] [doi: 10.1038/eye.2014.333] [Medline: 25592120]
82. Li Z, Liu F, Yang W, Peng S, Zhou J. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans Neural Netw Learn Syst 2021 Jun 10;PP (forthcoming). [doi: 10.1109/TNNLS.2021.3084827] [Medline: 34111009]
83. Breiman L. Random forests. Mach Learn 2001 Oct;45:5-32. [doi: 10.1023/A:1010933404324]
84. Bühren J, Kook D, Yoon G, Kohnen T. Detection of subclinical keratoconus by using corneal anterior and posterior surface aberrations and thickness spatial profiles. Invest Ophthalmol Vis Sci 2010 Jul;51(7):3424-3432. [doi: 10.1167/iovs.09-4960] [Medline: 20164452]
85. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
86. Scruggs BA, Chan RV, Kalpathy-Cramer J, Chiang MF, Campbell JP. Artificial intelligence in retinopathy of prematurity diagnosis. Transl Vis Sci Technol 2020 Feb 10;9(2):5 [FREE Full text] [doi: 10.1167/tvst.9.2.5] [Medline: 32704411]
87. Yim J, Chopra R, Spitz T, Winkens J, Obika A, Kelly C, et al. Predicting conversion to wet age-related macular degeneration using deep learning. Nat Med 2020 Jun;26(6):892-899. [doi: 10.1038/s41591-020-0867-7] [Medline: 32424211]
88. Chowdhury K, Dore C, Burr JM, Bunce C, Raynor M, Edwards M, et al. A randomised, controlled, observer-masked trial of corneal cross-linking for progressive keratoconus in children: the KERALINK protocol. BMJ Open 2019 Sep 12;9(9):e028761 [FREE Full text] [doi: 10.1136/bmjopen-2018-028761] [Medline: 31515418]
89. Larkin DF, Chowdhury K, Burr JM, Raynor M, Edwards M, Tuft SJ, KERALINK Trial Study Group. Effect of corneal cross-linking versus standard care on keratoconus progression in young patients: the KERALINK randomized controlled trial. Ophthalmology 2021 Nov;128(11):1516-1526 [FREE Full text] [doi: 10.1016/j.ophtha.2021.04.019] [Medline: 33892046]
90. Umemneku Chikere CM, Wilson K, Graziadio S, Vale L, Allen AJ. Diagnostic test evaluation methodology: a systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard - An update. PLoS One 2019 Oct;14(10):e0223832 [FREE Full text] [doi: 10.1371/journal.pone.0223832] [Medline: 31603953]
91. Flynn TH, Sharma DP, Bunce C, Wilkins MR. Differential precision of corneal Pentacam HR measurements in early and advanced keratoconus. Br J Ophthalmol 2016 Sep;100(9):1183-1187. [doi: 10.1136/bjophthalmol-2015-307201] [Medline: 26659714]
92. Yang K, Xu L, Fan Q, Zhao D, Ren S. Repeatability and comparison of new Corvis ST parameters in normal and keratoconus eyes. Sci Rep 2019 Oct 25;9(1):15379 [FREE Full text] [doi: 10.1038/s41598-019-51502-4] [Medline: 31653884]
93. Ahn SJ, Kim MK, Wee WR. Topographic progression of keratoconus in the Korean population. Korean J Ophthalmol 2013 Jun;27(3):162-166 [FREE Full text] [doi: 10.3341/kjo.2013.27.3.162] [Medline: 23730107]
94. Duncan JK, Belin MW, Borgstrom M. Assessing progression of keratoconus: novel tomographic determinants. Eye Vis (Lond) 2016 Mar 11;3:6 [FREE Full text] [doi: 10.1186/s40662-016-0038-6] [Medline: 26973847]
95. Meyer JJ, Gokul A, Vellara HR, Prime Z, McGhee CN. Repeatability and Agreement of Orbscan II, Pentacam HR, and Galilei Tomography Systems in Corneas with keratoconus. Am J Ophthalmol 2017 Mar;175:122-128. [doi: 10.1016/j.ajo.2016.12.003] [Medline: 27993593]
96. Guilbert E, Saad A, Elluard M, Grise-Dulac A, Rouger H, Gatinel D. Repeatability of keratometry measurements obtained with three topographers in keratoconic and normal corneas. J Refract Surg 2016 Mar;32(3):187-192. [doi: 10.3928/1081597X-20160113-01] [Medline: 27027626]

97. Hashemi H, Yekta A, Khabazkhoob M. Effect of keratoconus grades on repeatability of keratometry readings: comparison of 5 devices. J Cataract Refract Surg 2015 May;41(5):1065-1072. [doi: 10.1016/j.jcrs.2014.08.043] [Medline: 26049838]
98. Kanellopoulos AJ, Moustou V, Asimellis G. Evaluation of visual acuity, pachymetry and anterior-surface irregularity in keratoconus and crosslinking intervention follow-up in 737 cases. Int J Kerat Ect Cor Dis 2013:95-103. [doi: 10.5005/jp-journals-10025-1060]
99. Kanellopoulos AJ, Asimellis G. Revisiting keratoconus diagnosis and progression classification based on evaluation of corneal asymmetry indices, derived from Scheimpflug imaging in keratoconic and suspect cases. Clin Ophthalmol 2013;7:1539-1548 [FREE Full text] [doi: 10.2147/OPTH.S44741] [Medline: 23935360]
100. Piñero DP, Alio JL, Tomás J, Maldonado MJ, Teus MA, Barraquer RI. Vector analysis of evolutive corneal astigmatic changes in keratoconus. Invest Ophthalmol Vis Sci 2011 Jun 08;52(7):4054-4062. [doi: 10.1167/iovs.10-6856] [Medline: 21372010]
101. Yousefi S, Kiwaki T, Zheng Y, Sugiura H, Asaoka R, Murata H, et al. Detection of longitudinal visual field progression in glaucoma using machine learning. Am J Ophthalmol 2018 Sep;193:71-79. [doi: 10.1016/j.ajo.2018.06.007] [Medline: 29920226]
102. Belghith A, Bowd C, Medeiros FA, Balasubramanian M, Weinreb RN, Zangwill LM. Glaucoma progression detection using nonlocal Markov random field prior. J Med Imaging (Bellingham) 2014 Oct;1(3):034504 [FREE Full text] [doi: 10.1117/1.JMI.1.3.034504] [Medline: 26158069]
103. Hardcastle AJ, Liskova P, Bykhovskaya Y, McComish BJ, Davidson AE, Inglehearn CF, et al. A multi-ethnic genome-wide association study implicates collagen matrix integrity and cell differentiation pathways in keratoconus. Communications Biology 2021 Mar 01;4(1):266 [FREE Full text] [doi: 10.1038/s42003-021-01784-0] [Medline: 33649486]

Abbreviations

AS-OCT: anterior segment optical coherence tomography
AUC: area under the receiver operating characteristic curve
BAD-D: Belin/Ambrosio enhanced ectasia display
BFS: best-fit sphere
CNN: convolutional neural network
CXL: corneal collagen cross-linking
ESI: ectasia screening index
FFKC: forme fruste keratoconus
HOA: higher-order aberration
LASIK: laser-assisted in situ keratomileusis
NN: neural network
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS: Quality Assessment of Diagnostic Accuracy Studies
SVM: support vector machine

Edited by R Kukafka, G Eysenbach; submitted 25.01.21; peer-reviewed by A Chatterjee, RS Mahmoud; comments to author 06.04.21; revised version received 10.05.21; accepted 14.10.21; published 13.12.21.

Please cite as:
Maile H, Li JPO, Gore D, Leucci M, Mulholland P, Hau S, Szabo A, Moghul I, Balaskas K, Fujinami K, Hysi P, Davidson A, Liskova P, Hardcastle A, Tuft S, Pontikos N
Machine Learning Algorithms to Detect Subclinical Keratoconus: Systematic Review
JMIR Med Inform 2021;9(12):e27363
URL: https://medinform.jmir.org/2021/12/e27363
doi:10.2196/27363
PMID:34898463

Review

Artificial Intelligence in Predicting Cardiac Arrest: Scoping Review

Asma Alamgir[1*], BSc; Osama Mousa[1*], BSc; Zubair Shah[1,2], PhD

[1]College of Science and Engineering, Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

[2]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia

[*]these authors contributed equally

Corresponding Author:
Zubair Shah, PhD
College of Science and Engineering
Hamad Bin Khalifa University
Qatar Foundation
Education City, PO BOX 34110
Street 2731, Al Luqta St, Ar-Rayyan
Doha
Qatar
Phone: 974 5074 4851
Email: zshah@hbku.edu.qa

Abstract

Background: Cardiac arrest is a life-threatening cessation of activity in the heart. Early prediction of cardiac arrest is important, as it allows for the necessary measures to be taken to prevent or intervene during the onset. Artificial intelligence (AI) technologies and big data have been increasingly used to enhance the ability to predict and prepare for the patients at risk.

Objective: This study aims to explore the use of AI technology in predicting cardiac arrest as reported in the literature.

Methods: A scoping review was conducted in line with the guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for scoping reviews. Scopus, ScienceDirect, Embase, the Institute of Electrical and Electronics Engineers, and Google Scholar were searched to identify relevant studies. Backward reference list checks of the included studies were also conducted. Study selection and data extraction were independently conducted by 2 reviewers. Data extracted from the included studies were synthesized narratively.

Results: Out of 697 citations retrieved, 41 studies were included in the review, and 6 were added after backward citation checking. The included studies reported the use of AI in the prediction of cardiac arrest. Of the 47 studies, we were able to classify the approaches taken by the studies into 3 different categories: 26 (55%) studies predicted cardiac arrest by analyzing specific parameters or variables of the patients, whereas 16 (34%) studies developed an AI-based warning system. The remaining 11% (5/47) of studies focused on distinguishing patients at high risk of cardiac arrest from patients who were not at risk. Two studies focused on the pediatric population, and the rest focused on adults (45/47, 96%). Most of the studies used data sets with a size of <10,000 samples (32/47, 68%). Machine learning models were the most prominent branch of AI used in the prediction of cardiac arrest in the studies (38/47, 81%), and the most used algorithm was the neural network (23/47, 49%). K-fold cross-validation was the most used algorithm evaluation tool reported in the studies (24/47, 51%).

Conclusions: AI is extensively used to predict cardiac arrest in different patient settings. Technology is expected to play an integral role in improving cardiac medicine. There is a need for more reviews to learn the obstacles to the implementation of AI technologies in clinical settings. Moreover, research focusing on how to best provide clinicians with support to understand, adapt, and implement this technology in their practice is also necessary.

(JMIR Med Inform 2021;9(12):e30798) doi:10.2196/30798

KEYWORDS

artificial intelligence; machine learning; deep learning; cardiac arrest; predict

Introduction

Background

Cardiac arrest, also known as sudden cardiac death, is the cessation of the ability of the heart to pump blood. This acute cessation requires immediate intervention, as vital organs, such as the brain and the heart itself, are deprived of blood flow. A delay in intervention can lead to lifelong complications and even death. The global rate of mortality after cardiac arrest is significantly high—78% of out-of-hospital cardiac arrest (OHCA) cases die before they reach the hospital [1]. For those who do receive advanced care, the survival rate remains low. The survival rate for OHCA from the time of cardiac arrest to the time of discharge ranges from 2% to 11% worldwide [2]. The number of cardiac arrest deaths that occur within an in-hospital setting is also significant. In the United States alone, over 290,000 in-hospital cardiac arrests occur annually, with survival rates varying from as low as 0% to 36.2%, out of which a small percentage have favorable neurological prognoses [3].

Artificial intelligence (AI) is reforming health care every day. AI technologies have the perfect platform to thrive and mature with the growing adoption of electronic health records, development in computational power, continuous monitoring systems, and availability of big data [4]. It has become an important clinical decision-making tool that allows for personalized diagnoses, solutions, prognoses, and predictions of future health outcomes, guiding clinicians and other stakeholders in doing what is best for their patients [4]. AI technology is also rapidly progressing in cardiology, like in any other field of medicine [5]. AI-guided diagnosis and therapy selection have allowed for advancement in research, clinical practice, and population health in cardiovascular medicine [6]. Machine learning (ML) models have also been shown to outperform traditional statistical models in detecting sex differences in cardiovascular disease, further enhancing individualized medicine [7]. AI also plays a major role in improving care for cardiac arrest. AI technologies are being used to prevent cardiac arrest through early identification of risk factors [8], early detection [9], improved management (eg, effective cardiopulmonary resuscitation) [10], and prognosis determination for patients post cardiac arrest [11]. A large part of cardiac arrest research is the prediction of cardiac arrest before its occurrence, as it gives clinicians time to prepare and achieve better patient outcomes.

Thus, what are AI technologies and their counterparts in this context? AI refers to the field of science revolving around building computational systems and algorithms that facilitate the ability of a machine to mimic human behavior to learn and find solutions to tasks autonomously [4,12]. ML is a subset of AI. ML algorithms focus on building smart solutions after learning from patterns and experiences provided by a structured sample of training data [12]. Deep learning (DL) is a class of ML. It consists of a complex, interconnected, multilayered neural network, resembling a human brain. The aim of DL is to learn and understand patterns from a large amount of unstructured data [5]. In short, the more information it is fed, the more accurate the outcome.

The ability of AI technologies to process and evaluate patient data to generate predictions is important to support clinicians in making critical decisions, provide effective management, and, ultimately, improve patient outcomes in cardiac arrest cases [13]. Therefore, we believe it is crucial to explore the use of AI technology in predicting cardiac arrest and report our findings to help clinicians and researchers.

Research Problem and Aim

Numerous studies have proposed the use of AI in cardiac care, especially the use of AI in the prediction of cardiac arrest. However, there is a lack of consolidating existing evidence that describes the features of AI technologies, data sets, and data sources currently being used. It is essential to summarize recent findings that allow health care providers and researchers to implement appropriate guidelines, as well as to identify research gaps in the current literature. We encountered one review that examined the use of AI in the prediction of cardiac arrest [14]. However, the review was conducted in 2018 and did not include a large influx of studies in the past 2 years. Therefore, it is necessary to conduct a scoping review that focuses on various types of AI technologies currently being used in different settings to predict cardiac arrest.

This scoping review aims to explore the use and features of AI technologies applied to the prediction of cardiac arrest as reported in the literature. The results of our review will be a useful reference for health care professionals, researchers, and others involved in patient care to understand the application of AI and leverage it for the benefit of the community.

Methods

The scoping review was conducted by AA and OM to address this objective. The guidelines of the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) extension for scoping reviews [15] were followed to help conduct a transparent review.

Search Strategy

Five bibliographic databases were searched for this study: Scopus, ScienceDirect, Embase, the Institute of Electrical and Electronics Engineers, and Google Scholar. The databases were searched using search terms related to the target technology, population, and outcomes of interest. Search terms for our population included *Cardiac Arrest* OR *Heart Arrest* OR *Sudden Cardiac Death* OR *asystole* OR *cardiopulmonary arrest* and, for our intervention, *Artificial Intelligence* OR *Deep Learning* OR *Machine Learning* OR *Natural Language Processing* OR *Neural network* OR *Supervised learning* OR *Unsupervised learning* OR *Data mining*. Outcome- or purpose-related search terms included *Detect** OR *Predict** OR *Anticipat** OR *Diagnos**. The search query used for each database is presented in Multimedia Appendix 1.

For ScienceDirect and Google Scholar, only the first 100 and 50 results, respectively, were considered. This is because the reviewers found that the results became less relevant to the topic of interest and applicability after the mentioned number of citations. In addition to searching the databases, a backward reference list screening of the included studies was also carried

out to identify additional relevant studies. The search was conducted between March 15 and 20, 2021.

Eligibility Criteria

AI technologies implemented to predict cardiac arrest were included, with no restrictions on age, gender, geography, and type of AI technology used. Studies that focused primarily on predicting cardiac arrest were included. In contrast, studies dedicated to other aspects or contributing factors of cardiac arrest, such as arrhythmia and other cardiac diseases, were excluded. The review included peer-reviewed articles, preprints, articles in press, conference proceedings, theses, and dissertations written in English. Reviews, conference abstracts, study protocols, and proposals were excluded. No restrictions were imposed on the study design, study setting, country of publication, and publication year during the search query. However, only studies published between 2013 and 2021 were included in the review. The period between 2013 and 2016 constitutes a time when AI technologies saw a rapid increase of 175% in application [16]; therefore, the reviewers considered it to be a reasonable time period to include. The study eligibility criteria are summarized in Textbox 1.

Textbox 1. Inclusion and exclusion criteria.

Inclusion criteria

- Studies that focused on the use of artificial intelligence (AI) technologies in cardiac arrest prediction for the benefit of the human population
- Studies published from 2013 to 2021
- Peer-reviewed articles, articles in press, theses, dissertations, and conference proceedings
- Primary studies

Exclusion criteria

- Articles that did not address the use of AI in cardiac arrest prediction
- Reviews, conference abstracts or proposals, letters, news, books, and protocols
- Published in a language other than English

Study Selection

The studies retrieved from the databases were first imported to Rayyan (Rayyan System Inc) [17], a collaborative research tool, to undergo 3 phases of the filtering process. This ensured that the articles we included in the review were relevant to our study objective. The 3 phases of the filtering process were as follows: (1) identification phase, where citations were identified after applying the search terms to the databases and duplicates were removed; (2) screening phase, where titles and abstracts were screened to remove articles that did not match our inclusion criteria; and (3) eligibility phase, where the full texts of the articles were read to determine their applicability on the basis of the inclusion criteria. The 2 reviewers conducted all 3 phases independently, facilitated by the Rayyan application. In case of conflict, a discussion was held to reach a consensus.

Data Extraction and Data Synthesis

To conduct a reliable and consistent extraction of data from the included studies, a data extraction form was used (Multimedia Appendix 2). The 2 reviewers independently extracted data related to the characteristics of the included studies, AI technology, and data sets. The extracted information was recorded on a shared Microsoft Excel sheet for easy data management. Similar to the study selection, any conflict between the 2 reviewers was resolved through discussions to reach a consensus.

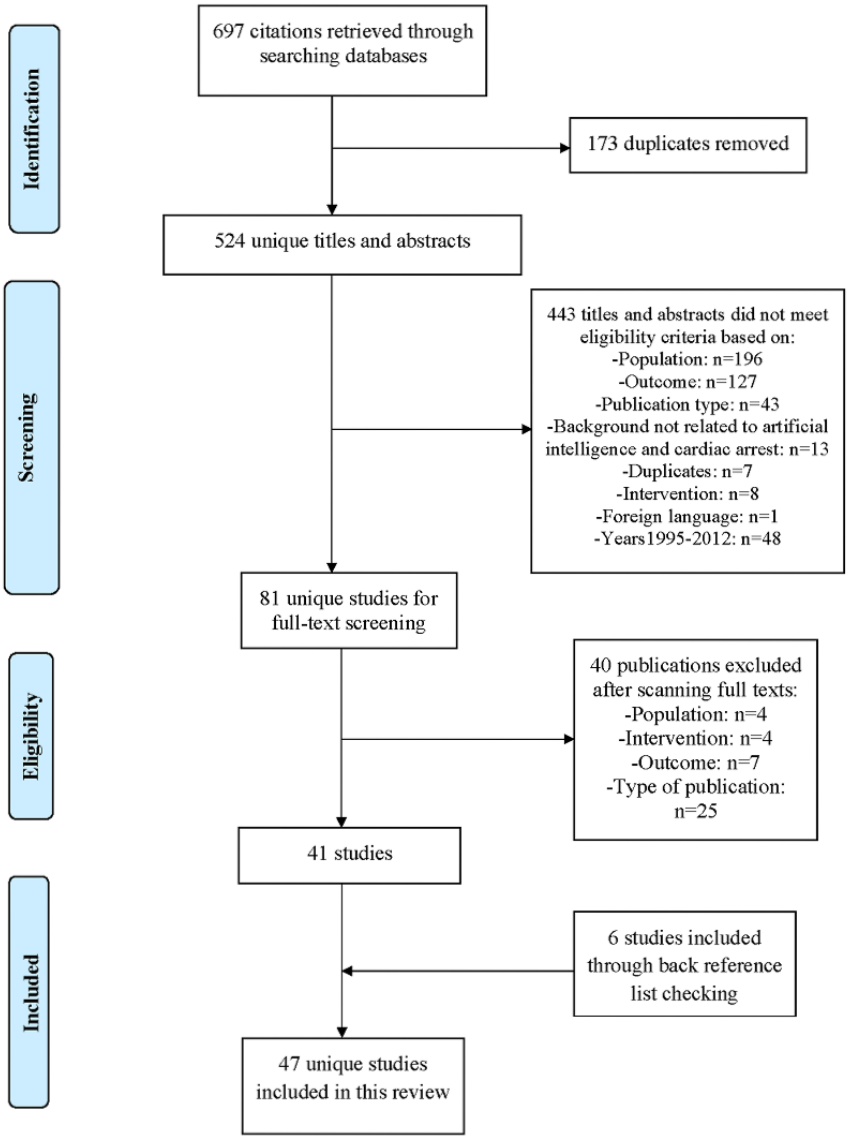
A narrative synthesis of the extracted data was performed. The findings from the included studies were classified and described in terms of their purpose, AI branch, algorithm, and platform used to implement the algorithm. The data sets used for the development and validation of the technology were considered and described. The data sources, size of the data set. validation type, and proportion of training, validation, and test data sets were included when available. An sheet Excel (Multimedia Appendix 3) was used to record the extracted data to facilitate data synthesis.

Results

Search Findings

As shown in Figure 1, 697 studies were retrieved from our search, of which 173 (24.8%) duplicates were removed. A total of 524 underwent title and abstract screening, of which 443 (84.5%) studies were excluded. The reasons for exclusion are shown in Figure 1. In total, 81 unique studies underwent full-text screening to evaluate eligibility, of which 41 (51%) studies met the inclusion criteria and were included in the review. Six additional studies were identified and added by checking the reference lists of those 41 studies. Overall, 47 studies were included in the review.

Figure 1. Flowchart of the study selection process.



Characteristics of the Included Studies

Of the 47 studies included, 46 (98%) were published in peer-reviewed journals, whereas 1 (2%) was still in press. Approximately 81% (38/47) of the studies were research articles, whereas the rest were conference proceedings (9/47, 19%). Only 2 studies from 2013 were included, whereas most of the studies were from 2020 (12/47, 26%). The other included studies were conducted in 2014 (4/47, 9%), 2015 (5/47, 11%), 2016 (3/47, 6%), 2017 (3/47, 6%), 2018 (5/47, 11%), 2019 (9/47, 19%), and 2021 (4/47, 9%). The included studies were conducted in 15 countries, and most of the studies were published in India and the United States (9/47, 19%). Table 1 shows the characteristics of the studies included in our review. Multimedia Appendix 3 demonstrates the attributes of each study.

Table 1. Characteristics of the included studies (N=47).

Characteristic	Studies, n (%)
Paper status	
Published	46 (98)
In press	1 (2)
Publication type	
Conference proceeding	9 (19)
Research article	38 (81)
Country	
Australia	1 (2)
China	3 (6)
Greece	1 (2)
India	9 (19)
Iran	5 (11)
Japan	1 (2)
Malaysia	4 (9)
Poland	1 (2)
Portugal	1 (2)
Singapore	1 (2)
South Korea	7 (15)
Spain	1 (2)
Taiwan	2 (4)
United Kingdom	1 (2)
United States	9 (19)
Year published	
2013	2 (4)
2014	4 (9)
2015	5 (11)
2016	3 (6)
2017	3 (6)
2018	5 (11)
2019	9 (19)
2020	12 (26)
2021	4 (9)

AI Characteristics in the Included Studies

Use of AI in Predicting Cardiac Arrest

The approaches taken by the included studies to predict cardiac arrest using AI technologies were divided into 3 categories: analysis of variables and parameters, development of an early warning system or prediction model, and stratification of patients at a high risk of cardiac arrest.

Analysis of Variables and Parameters

The studies in this category focused on analyzing one or more patient parameters to determine their impact on the efficiency of improving the prediction of cardiac arrest in combination with AI algorithms. We observed 26 studies that fit into this category [14,18-42]. Of these 26 studies, 11 (42%) used ML models [14,18,19,23,25,27,30,32,34-36] and 3 (12%) used DL algorithms [20,31,38]. We observed that 12 studies incorporated both ML and DL models to analyze and validate different parameters [21,22,24,26,28,29,33,37,39-42].

Random forest (RF) [14,21,23,28-30,32,35-37,39-41] and support vector machine (SVM) [18,22,24,26,28,34,40-42] were the most used ML models observed in these studies, followed by decision tree (DT) [22,29,30,40-42], logistic regression (LR) [28-30,40], Naive Bayes [19,28,29,41], gradient boosting

[27,28], extreme gradient boosting [27,29], LogitBoost [21], AdaBoost [29], TreeBagger [34], and sequential feature selection [24]. The most used DL-based algorithm in the studies was k-nearest neighbors (KNN) [20,22,26,29,33,41,42]. The probabilistic neural network [24,31,42], artificial neural network [29,40], multilayer perceptron [21,33], long short-term memory [39], convolutional neural network [37], and enhanced probabilistic neural network [31] were also algorithms used in DL model studies. Furthermore, 2 studies did not specify the algorithm used [25,38].

The parameters analyzed and validated in the included studies were diverse. The majority of the studies focused on using various characteristics from patient electrocardiogram readings [14,18,20-22,24,26,27,30-34,36-38,40-42], especially heart rate variability (HRV) [14,21,22,26,30,32,34,36-38,42]. HRV is the variation in time between each heartbeat that can be tracked on an electrocardiogram [43]. This noninvasive assessment tool provides important information about the autonomic nervous system, allowing clinicians to determine current and impending cardiac disease [44]. Its usefulness in determining cardiac-related prognosis is also well-documented in the literature [45,46]. In the included studies, HRV appeared to improve prediction outcomes in the studies that integrated it into the data set. All studies using HRV reported higher performance in terms of accuracy and other outcome indicators. Other unique parameters, such as genetic data [20], smoking habit [29], nursing documentation [25], and dialysis status [23,28], were also used to evaluate their effect on the performance of the AI technology to predict cardiac arrest. Accuracy [18,19,21,22,29,31,33,34,36,37,40,41] and sensitivity [14,22,26-28,30,32-34,41,42] were the most used measures of outcome in this category.

Development of an Early Warning System Using AI

In 16 studies [47-62], the focus of AI technologies was to develop an early warning system alerting health care professionals when patients were at risk of going into cardiac arrest in the future. To develop a warning model, most studies used ML model algorithms [49-51,53,54,59,61], whereas 5 only used DL-based algorithms [47,48,56,60,62]. Four studies used both ML- and DL-based algorithms [52,55,57,58], comparing them with each other to observe which yielded the best outcome. The ML algorithms used in these studies included LR [50,52,55,58], SVM [50-52,58], DT [52,53,57,59], RF [55,57,58], Naive Bayes [57,58], gradient boosting [58], Bayesian networks [49], AdaBoost [57], transfer learning [54], and multichannel Hidden Markov Model [61]. The KNN [52,58], artificial neural network [48,58,61], long short-term memory [47,56,59], and recurrent neural network [47,55,56,62] algorithms were used in the studies to constitute a DL-based early warning system. A total of 10 of the studies compared their outcomes to existing or *traditional* early warning systems [47,48,50,52,53,56-58,60,62]. The studies compared their models to scoring systems such as the Modified Early Warning Score [48,50,52,53,56,58,60,62], Early Warning Score [57], National Early Warning Score [60], and Pediatric Early Warning Score [47]. Only 1 study showed similar outcomes when using an AI model compared with a traditional warning system [53], whereas, in other studies, the AI-based model outperformed the system it was compared with. For example, deep early warning systems detected 50%-78% more cardiac arrests compared with the Modified Early Warning Score [56,62]. Moreover, the prediction period of the algorithms was reported to range from 30 minutes to as early as 24 hours before the onset of cardiac arrest [50,53,57,58,62].

Three of the most used outcome measures in this category included the area under the receiver operating characteristic curve [47,48,51,53,55-57,60,62], sensitivity [49,52,58,60,62], and accuracy [51,54,58-60].

Stratification of High-risk Patients

In 5 studies [63-67], AI technologies were used to distinguish patients who were at high risk of cardiac arrest from patients who were not at risk. Three studies highlighted HRV [63-65] as an important feature to distinguish high-risk patients.

ML was used in the majority of the studies [63,64,67], and only 1 study used a DL algorithm [66]. One study used both ML and DL models to stratify patients [65]. The ML algorithms used were SVM [63,64], linear discriminant analysis [64], DT [63], LR [67], RF [67], extreme gradient boosting [67], and fuzzy classifier [65]. The DL algorithms included KNN [65,66] and multilayer perceptron [66]. The outcome measures in the studies included accuracy [63-66], sensitivity, specificity [63-65], area under the receiver operating characteristic curve, and the precision-recall curve [66].

Features of AI Techniques in the Studies

Most studies used traditional ML models and algorithms to predict cardiac arrest (38/47, 81%) whereas 55% (26/47) used DL techniques. We observed 15 types of AI classifiers used in the studies to predict cardiac arrest **(**Table 2**)**. A notable observation is that 6 models were commonly used; neural network–based models, which are a DL model, and RF, which is a traditional ML model, were used 20 and 18 times, respectively, making them the top 2 most used models found in the studies, followed by SVM (15/47, 32%), DT (12/47, 26%), LR (11/47, 23%), and KNN (10/47, 21%). Less common models, such as transfer learning, linear discriminant analysis, fuzzy classifier, multichannel Hidden Markov Model, LogitBoost, AdaBoost, Bayesian networks, Naive Bayes, and extreme gradient boosting, were used between 1 and 6 times in the studies. Two studies used wearable devices as the platform for their AI techniques [24,59], whereas the remaining studies used computers. Multimedia Appendix 3 presents the features of the AI techniques.

Table 2. Features of artificial intelligence (AI)–based techniques used for cardiac arrest prediction (N=47).

Feature	Study ID[a]	Studies, n (%)[b]
AI model[c]		
Neural network	1, 3, 4, 6, 11, 13, 14, 15, 16, 19, 21, 25, 26, 28, 32, 34, 26, 38, 45, 46	20 (43)
Random forest	3, 6, 7, 8, 9, 10, 13, 14, 15, 17, 18, 19, 28, 20, 35, 37, 41, 45	18 (38)
Support vector machine	2, 5, 19, 20, 27, 30, 31, 32, 34, 38, 41, 42, 43, 45, 46	15 (32)
Decision tree	3, 5, 15, 16, 17, 18, 19, 20, 32, 34, 40, 42	12 (26)
Logistic regression	3, 6, 10, 15, 16, 18, 19, 30, 32, 45, 47	11 (23)
K-nearest neighbors	3, 20, 24, 32, 33, 34, 36, 42, 43, 46	10 (21)
Extreme gradient boosting	3, 10, 15, 16, 44, 45	6 (13)
Naive Bayes	16, 20, 22, 45	4 (9)
AdaBoost	15	1 (2)
Bayesian networks	29	1 (2)
LogitBoost	28	1 (2)
Multichannel Hidden Markov Model	23	1 (2)
Fuzzy classifier	33	1 (2)
Linear discriminant analysis	27	1 (2)
Transfer learning	47	1 (2)
Platform		
Computer	1-16, 18-37, 39-47	45 (96)
Wearable	17, 38	2 (4)

[a]The order of the reviewed studies in this table follows the order shown in Multimedia Appendix 3.

[b]Two studies did not specify the artificial intelligence model used.

[c]The numbers do not add up as some studies used more than one artificial intelligence model or algorithm.

Features of Data Sets Used for Development and Validation of AI Models

Clinical setting sources (such as hospital databases and medical centers) were the most commonly used data sources for the development and validation of AI models [14,25,27,28,31,32,34-36,38,39,47-53,55-57,60,62,67]. Public resources (eg, the MIT-BIH Arrhythmia and Normal Sinus Rhythm databases) [18-24,26,29,30,33,37,41,42,54,58,61,63-66] were the other sources of data for AI models.

Several types of data were retrieved from these sources. We grouped the types of data into 5 categories: clinical data, demographic data (eg, age, gender, and ethnicity), laboratory data (eg, blood samples), radiology data (eg, x-rays), and biological data (eg, genetic information). As shown in Table 3, 58% (34/47) of the studies used clinical data as the data type. Different variables fall under this category; Table 4 breaks down the type of clinical data observed in the studies. Demographic data were the second most used data type in predicting cardiac arrest (15/47, 26%), followed by laboratory data (8/47, 14%) and biological data (1/47, 2%).

Table 3. Data types.

Data type	Studies, n (%)
Clinical data	34 (72)
Demographic data	15 (32)
Laboratory data	8 (17)
Biological data	1 (2)

Table 4. Clinical data breakdown[a].

Clinical data types	Studies, n (%)
Vital signs	23 (49)
ECG[b] variables	18 (38)
Medical history	10 (21)
Chief complaint	3 (6)
Medication	3 (6)
Cardiopulmonary exercise testing	2 (4)
Diagnosis	2 (4)
Risk score	2 (4)
Renal status	2 (4)
Cardiopulmonary resuscitation information	1 (2)
Lifestyle	1 (2)
Nursing notes	1 (2)

[a]Several studies collected more than one clinical data type.

[b]ECG: echocardiogram.

For data set sizes, 42 (89%) out of 47 studies mentioned the size of the training data set used for the ML model. Of the 47 studies, 23 (49%) used data sets of less than 1000 samples, whereas 14 (30%) used data sets of between 1000 and 9999 samples. Moreover, 11% (5/47) of studies used more than 10,000 data samples. Various validation types for the AI models were reported in 41 studies. These validation methods were divided into 3 main categories: k-fold cross-validation, which was the most common validation technique used (24/47, 51%), followed by train-test split (11/47, 23%) and external validation (6/47, 13%). Table 5 provides a breakdown of the features of data used in the included studies.

Table 5. Features of the data used (N=47).

Feature	Studies, n (%)
Data sources	
Public database	21 (45)
Clinical setting	24 (51)
Other	2 (4)
Data set size[a]	
<1000	23 (49)
1000-9999	14 (28)
≥10,000	5 (11)
Type of validation[b]	
K-fold cross-validation	24 (51)
Train-test split	11 (23)
External validation	6 (13)

[a]Data set size mentioned in 42 studies.

[b]Types of validation mentioned in only 41 studies.

Discussion

Principal Findings

In this review, we explored the use of AI in predicting cardiac arrest. From a total of 617 retrieved studies, 47 (7.6%) were included in this review. We found that the number of studies increased in the past 2 years (9 in 2019 and 11 in 2020), which is not surprising given that the use of AI technology in health care has been increasing. India and the United States (9/47, 19%) represent the countries that published the most studies related to AI in predicting cardiac arrest, with a total of 18. To explore the use of AI technology in predicting cardiac arrest, we divided our findings into 3 categories, each representing a classification of the reviewed studies from a different perspective. The first category focuses on the way AI

technologies are used in predicting cardiac arrest and comprises 3 main subcategories: (1) stratification between patients with cardiac arrest and non-at-risk patients, in which the AI technology was trained using the history of patients who had cardiac arrest and classified patients with a high risk of cardiac arrest; (2) development of an early warning system using AI, in which AI technology was used to alert physicians 1 to 16 hours before cardiac arrest and its accuracy was compared with other existing traditional warning systems; and (3) analysis of different variables and parameters to observe the efficiency of prediction.

The second category identifies the features of the AI techniques as observed in the literature. Two AI branches were used, ML and DL, where ML was the most used branch in a total of 38 studies, and the most used model in this branch was RF (18/47, 38%). In contrast, DL was used 16 times, and the most used model were neural network–based models (20/47, 43%). Finally, the third category classifies the data and validation method used for the AI, where we expanded on the data sources, data types, and validation processes found in the literature for the AI techniques. A total of 42 out of the 47 studies mentioned the data set size used, the majority of the studies using data sets of less than 1000 samples (23/47, 49%). Most studies used k-fold cross-validation to test the AI models (24/47, 51%).

The Implications for Practice and Research

This review highlighted the most common AI models used in predicting cardiac arrest and the different approaches used in predicting it. On the basis of our findings AI models can predict cardiac arrest using a variety of data types. In our review, ML techniques were used much more than DL techniques. One explanation for this is that the data used to train the AI model were mostly structured (eg, vital signs are recorded, and the threshold for the measurements of a normal human being is known and then compared with the vital signs of a patient who had cardiac arrest). Therefore, it is understandable that most researchers used ML techniques, because they were dealing with structured data. In contrast, DL works best with unstructured data, which was less commonly used in the articles reviewed. Another explanation is the size of the data sets used, as most studies used relatively small data sets to train DL models (eg, only 5 studies out of 47 used data sets of more than 10,000 samples). Finally, many studies explained the use of ML techniques such as DTs, LR, and RF, which consist of many DTs given that the main outcome is binary (at risk of cardiac arrest or not at risk of cardiac arrest). This explains the rapid use of these techniques in the reviewed studies.

Future research should explore ways to attain higher prediction accuracy in terms of the time before cardiac arrest may occur to the patient and the percentage of true positive and true negative (accurately predicting that the patient will experience cardiac arrest). Moreover, more research is required to address and investigate hyperparameter optimization, as it could lead to different performance results of ML models across the studies selected and influence which parameters are important for the prediction of cardiac arrest. Early prediction of cardiac arrest could be achieved through the correlation between the clinical data obtained and the demographic data of the patient. ML seems to be the best technique to be used because the data used is structured (eg, age, vital signs, and electrocardiogram variables). The earlier the prediction time, the higher the likelihood that the physicians can save the patients from sudden cardiac death. Furthermore, the potential to evaluate the effectiveness of less frequently used data types, such as laboratory and biological data, in predicting cardiac arrest should also be explored.

Only 5 studies reviewed used data sets of more than 10,000 samples, whereas most of the studies used data sets of less than 1000 samples. Future studies need to evaluate AI models using larger data sets to improve their effectiveness. In addition, comparing the prediction accuracy of AI techniques with each other is a good method of evaluation. However, AI techniques need to be compared with other techniques used to predict cardiac arrest.

Studies that did research in clinical settings limited the population to a specific hospital or country, which produced biased results that do not apply everywhere. Future studies should consider public databases that contain cases from different hospitals and countries.

Many studies explored the potential of AI in the prediction of arrhythmia and irregular heartbeat, and future studies should investigate the potential of the proposed models in the prediction of cardiac arrest. Finally, future research should explore the potential of physiological and psychological data in the prediction of cardiac arrest.

Strengths

The review addressed the use of all types of AI technologies to predict cardiac arrest in all populations with no restrictions on paper status, study settings, and geographic location in a comprehensive manner. Moreover, an in-depth exploration was conducted on the features of AI technology and the data sets that were used to develop and validate these technologies.

Other reviews have explored the use of ML and DL in detecting arrhythmia [53,68] or the use of AI in cardiology in general [69-71] but have not gone into detail on how this technology can be used to predict cardiac arrest. A previous systematic review explored the use of ML in predicting cardiac arrest [72]; however, to the best of our knowledge, this is the first review to explore the different approaches to predicting cardiac arrest to fill the research gap with a better understanding of the prediction techniques rather than focusing on whether the model was able to predict only cardiac arrest. Moreover, this study did not focus on a specific AI branch (ML, DL, or natural language); rather, it focused on categorizing the AI techniques into branches to provide insight into the most common AI technique in every branch.

The studies included in the review comprised the latest publications, reducing the selection bias date. In addition to published research articles, conference proceedings were also included to maximize the extent of inclusion. This was also done by conducting a backward reference list check of the included studies. Furthermore, study selection and data extraction involved 2 reviewers independently overseeing the process, which ensured minimal selection bias.

Limitations

This review did not include databases such as ACM and JSTOR, which limited our access to gray literature and other potentially relevant studies. This was because of the lack of access to some of the databases and others specialized in physiological or engineering studies rather than medical studies. Moreover, owing to practical constraints, only English-language studies were included in the review, excluding studies in other languages. Furthermore, our search query did not include MeSH (Medical Subject Headings) terms or algorithm-specific search terms, which might have hidden studies that would otherwise have been appropriate for our review.

Conclusions

Our scoping review included 47 studies that focused on the use of AI technologies to predict cardiac arrest in all settings. With the big data available from patient monitoring systems and electronic health records, it is possible to delve deeper into making our approach to cardiac arrest reliable and more effective, increasing the rate of survival over time. Moreover, with the increasing adoption of wearable devices with sensors tracking various aspects of health and activity, there are opportunities for research to develop techniques to predict and alert patients at risk of OHCAs. Furthermore, clinicians need to be on board with the rapidly growing technology as, without them, we cannot move forward. Therefore, more research on AI paired with education initiatives within health care professionals needs to be considered.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Search strategy.
[DOCX File , 14 KB - medinform_v9i12e30798_app1.docx]

Multimedia Appendix 2
Data extraction form.
[DOCX File , 14 KB - medinform_v9i12e30798_app2.docx]

Multimedia Appendix 3
Characteristics of the included studies and features of artificial intelligence techniques.
[DOCX File , 31 KB - medinform_v9i12e30798_app3.docx]

References

1. Yan S, Gan Y, Jiang N, Wang R, Chen Y, Luo Z, et al. The global survival rate among adult out-of-hospital cardiac arrest patients who received cardiopulmonary resuscitation: a systematic review and meta-analysis. Crit Care 2020 Feb 22;24(1):61 [FREE Full text] [doi: 10.1186/s13054-020-2773-2] [Medline: 32087741]
2. Berdowski J, Berg RA, Tijssen JG, Koster RW. Global incidences of out-of-hospital cardiac arrest and survival rates: systematic review of 67 prospective studies. Resuscitation 2010 Nov;81(11):1479-1487. [doi: 10.1016/j.resuscitation.2010.08.006] [Medline: 20828914]
3. Chan PS, Krein SL, Tang F, Iwashyna TJ, Harrod M, Kennedy M, American Heart Association's Get With the Guidelines–Resuscitation Investigators. Resuscitation practices associated with survival after in-hospital cardiac arrest: a nationwide survey. JAMA Cardiol 2016 May 01;1(2):189-197 [FREE Full text] [doi: 10.1001/jamacardio.2016.0073] [Medline: 27437890]
4. Bohr A, Memarzadeh K, editors. Artificial Intelligence in Healthcare. Amsterdam: Elsevier; 2020.
5. Lopez-Jimenez F, Attia Z, Arruda-Olson AM, Carter R, Chareonthaitawee P, Jouni H, et al. Artificial intelligence in cardiology: present and future. Mayo Clin Proc 2020 May;95(5):1015-1039. [doi: 10.1016/j.mayocp.2020.01.038] [Medline: 32370835]
6. Johnson KW, Soto JT, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018 Jun 12;71(23):2668-2679 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.521] [Medline: 29880128]
7. Sarajlic P, Plunde O, Franco-Cereceda A, Bäck M. Artificial intelligence models reveal sex-specific gene expression in aortic valve calcification. JACC Basic Transl Sci 2021 May;6(5):403-412 [FREE Full text] [doi: 10.1016/j.jacbts.2021.02.005] [Medline: 34095631]
8. JayaSree M, Koteswara Rao L. Survey on - identification of coronary artery disease using deep learning. Mater Today Proc 2020 Oct:526. [doi: 10.1016/j.matpr.2020.09.526]
9. Kumari CU, Murthy AS, Prasanna LW, Reddy MP, Panigrahy AK. An automated detection of heart arrhythmias using machine learning technique: SVM. Mater Today Proc 2021 Aug;45:1393-1398. [doi: 10.1016/j.matpr.2020.07.088]

10. Didon J, Ménétré S, Jekova I, Stoyanov T, Krasteva V. Analyze Whilst Compressing Algorithm for detection of ventricular fibrillation during CPR: a comparative performance evaluation for automated external defibrillators. Resuscitation 2021 Mar;160:94-102. [doi: 10.1016/j.resuscitation.2021.01.018] [Medline: 33524490]
11. Cronberg T, Greer DM, Lilja G, Moulaert V, Swindell P, Rossetti AO. Brain injury after cardiac arrest: from prognostication of comatose patients to rehabilitation. Lancet Neurol 2020 Jul;19(7):611-622. [doi: 10.1016/s1474-4422(20)30117-4]
12. de Marvao A, Dawes TJ, Howard JP, O'Regan DP. Artificial intelligence and the cardiologist: what you need to know for 2020. Heart 2020 Mar 23;106(5):399-400 [FREE Full text] [doi: 10.1136/heartjnl-2019-316033] [Medline: 31974212]
13. Miyazawa AA. Artificial intelligence: the future for cardiology. Heart 2019 Aug 12;105(15):1214. [doi: 10.1136/heartjnl-2018-314464] [Medline: 30636218]
14. Elola A, Aramendi E, Rueda E, Irusta U, Wang H, Idris A. Towards the prediction of rearrest during out-of-hospital cardiac arrest. Entropy (Basel) 2020 Jul 09;22(7):758 [FREE Full text] [doi: 10.3390/e22070758] [Medline: 33286529]
15. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med 2018 Oct 02;169(7):467-473 [FREE Full text] [doi: 10.7326/M18-0850] [Medline: 30178033]
16. Boom in artificial intelligence patents, points to ‘quantum leap’ in tech: UN report. UN News. 2019. URL: https://news.un.org/en/story/2019/01/1031702 [accessed 2021-04-08]
17. Rayyan QCRI. URL: https://www.rayyan.ai [accessed 2021-11-11]
18. Murugappan M, Murugesan L, Jerritta S, Adeli H. Sudden Cardiac Arrest (SCA) prediction using ECG morphological features. Arab J Sci Eng 2020 Jul 26;46(2):947-961. [doi: 10.1007/s13369-020-04765-3]
19. Karankar N, Shukla P, Agrawal N. Comparative study of various machine learning classifiers on medical data. In: Proceedings of the 7th International Conference on Communication Systems and Network Technologies (CSNT). 2017 Presented at: 7th International Conference on Communication Systems and Network Technologies (CSNT); Nov. 11-13, 2017; Nagpur, India p. 267-271. [doi: 10.1109/csnt.2017.8418550]
20. Alfarhan KA, Mashor MY, Zakaria A, Omar MI. Automated electrocardiogram signals based risk marker for early sudden cardiac death prediction. J Med Imaging Heal Informatics 2018 Dec 01;8(9):1769-1775. [doi: 10.1166/jmihi.2018.25311769]
21. Tapas N, Lone T, Reddy D, Kuppili V. Prediction of cardiac arrest recurrence using ensemble classifiers. Sādhanā 2017 Jun 17;42(7):1135-1141. [doi: 10.1007/s12046-017-0683-z]
22. Fujita H, Acharya UR, Sudarshan VK, Ghista DN, Sree SV, Eugene LW, et al. Sudden cardiac death (SCD) prediction based on nonlinear heart rate variability features and SCD index. Appl Soft Comput 2016 Jun;43:510-519. [doi: 10.1016/j.asoc.2016.02.049]
23. Goldstein BA, Chang TI, Mitani AA, Assimes TL, Winkelmayer WC. Near-term prediction of sudden cardiac death in older hemodialysis patients using electronic health records. Clin J Am Soc Nephrol 2013 Oct 31;9(1):82-91. [doi: 10.2215/cjn.03050313]
24. Murukesan L, Murugappan M, Iqbal M, Saravanan K. Machine Learning Approach for Sudden Cardiac Arrest Prediction Based on Optimal Heart Rate Variability Features. J Med Imaging Hlth Inform 2014 Aug 01;4(4):521-532. [doi: 10.1166/jmihi.2014.1287]
25. Collins SA, Cato K, Albers D, Scott K, Stetson PD, Bakken S, et al. Relationship between nursing documentation and patients' mortality. Am J Crit Care 2013 Jul 01;22(4):306-313 [FREE Full text] [doi: 10.4037/ajcc2013426] [Medline: 23817819]
26. Houshyarifar V, Chehel Amirani M. An approach to predict Sudden Cardiac Death (SCD) using time domain and bispectrum features from HRV signal. Biomed Mater Eng 2016 Aug 12;27(2-3):275-285. [doi: 10.3233/bme-161583]
27. Wu TT, Lin XQ, Mu Y, Li H, Guo YS. Machine learning for early prediction of in-hospital cardiac arrest in patients with acute coronary syndromes. Clin Cardiol 2021 Mar 14;44(3):349-356 [FREE Full text] [doi: 10.1002/clc.23541] [Medline: 33586214]
28. Nakajima K, Nakata T, Doi T, Tada H, Maruyama K. Machine learning-based risk model using I-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure. J Nucl Cardiol 2020 May 14;11(10):897-904 [FREE Full text] [doi: 10.1007/s12350-020-02173-6] [Medline: 32410060]
29. L PR, Jinny SV, Mate YV. Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques. Health Technol 2020 Nov 13;11(1):63-73. [doi: 10.1007/s12553-020-00508-4]
30. Shashikant R, Chetankumar P. Predictive model of cardiac arrest in smokers using machine learning technique based on Heart Rate Variability parameter. Appl Comput Informatics 2020 Jul 28:Ahead of Print. [doi: 10.1016/j.aci.2019.06.002]
31. Amezquita-Sanchez JP, Valtierra-Rodriguez M, Adeli H, Perez-Ramirez CA. A novel wavelet transform-homogeneity model for sudden cardiac death prediction using ECG signals. J Med Syst 2018 Aug 16;42(10):176. [doi: 10.1007/s10916-018-1031-5] [Medline: 30117048]
32. Liu N, Koh ZX, Goh J, Lin Z, Haaland B, Ting BP, et al. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. BMC Med Inform Decis Mak 2014 Aug 23;14:75 [FREE Full text] [doi: 10.1186/1472-6947-14-75] [Medline: 25150702]

33. Ebrahimzadeh E, Foroutan A, Shams M, Baradaran R, Rajabion L, Joulani M, et al. An optimal strategy for prediction of sudden cardiac death through a pioneering feature-selection approach from HRV signal. Comput Methods Programs Biomed 2019 Feb;169(3):19-36. [doi: 10.1016/j.cmpb.2018.12.001] [Medline: 30638589]
34. Mirhoseini SR, Jahedmotlagh MR, Pooyan M. Improve accuracy of early detection Sudden Cardiac Deaths (SCD) using decision forest and SVM. In: Proceedings of the International Conference on Robotics and Artificial Intelligence. 2016 Presented at: International Conference on Robotics and Artificial Intelligence; April 20-22, 2016; Los Angeles URL: https://www.researchgate.net/publication/296701627_Improve_Accuracy_of_Early_Detection_Sudden_Cardiac_Deaths_SCD_Using_Decision_Forest_and_SVM
35. Ueno R, Xu L, Uegami W, Matsui H, Okui J, Hayashi H, et al. Value of laboratory results in addition to vital signs in a machine learning algorithm to predict in-hospital cardiac arrest: a single-center retrospective cohort study. PLoS One 2020;15(7):e0235835 [FREE Full text] [doi: 10.1371/journal.pone.0235835] [Medline: 32658901]
36. Balachander T, Pradeep K, Balaji JS. An integrated approach for early risk detection of sudden cardiac death using machine learning approach. Int J Adv Sci Technol 2020;29(6):2500-2509 [FREE Full text]
37. RamKumar RP, Polepaka S. Performance comparison of random forest classifier and convolution neural network in predicting heart diseases. In: Advances in Intelligent Systems and Computing. Singapore: Springer; 2020:683-691.
38. Kwon JM, Kim KH, Jeon KH, Lee SY, Park JJ, Oh BH. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. Scand J Trauma Resusc Emerg Med 2020 Oct 06;28(1):98 [FREE Full text] [doi: 10.1186/s13049-020-00791-0] [Medline: 33023615]
39. Chang HK, Wu CT, Liu JH, Lim WS, Wang HC, Chiu SI, et al. Early detecting in-hospital cardiac arrest based on machine learning on imbalanced data. In: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI). 2019 Presented at: IEEE International Conference on Healthcare Informatics (ICHI); June 10-13, 2019; Xi'an, China. [doi: 10.1109/ICHI.2019.8904504]
40. Chauhan U, Kumar V, Chauhan V, Tiwary S, Kumar A. Cardiac Arrest Prediction using Machine Learning Algorithms. In: Proceedings of the 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). 2019 Presented at: 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT); 5-6 July 2019; Kannur, India. [doi: 10.1109/ICICICT46008.2019.8993296]
41. Lai D, Zhang Y, Zhang X, Su Y, Bin Heyat MB. An automated strategy for early risk identification of sudden cardiac death by using machine learning approach on measurable arrhythmic risk markers. IEEE Access 2019;7(2):94701-94716. [doi: 10.1109/access.2019.2925847]
42. Acharya UR, Fujita H, Sudarshan V, Ghista D, Eugene L, Koh J. Automated prediction of sudden cardiac death risk using kolmogorov complexity and recurrence quantification analysis features extracted from HRV signals. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics. 2019 Presented at: IEEE International Conference on Systems, Man, and Cybernetics; Oct. 9-12, 2015; Hong Kong, China p. 94701-94716. [doi: 10.1109/SMC.2015.199]
43. Acharya UR, Joseph KP, Kannathal N, Lim CM, Suri JS. Heart rate variability: a review. Med Biol Eng Comput 2006 Dec;44(12):1031-1051. [doi: 10.1007/s11517-006-0119-0] [Medline: 17111118]
44. ChuDuc H, NguyenPhan K, NguyenViet D. A review of heart rate variability and its applications. APCBEE Procedia 2013;7:80-85. [doi: 10.1016/j.apcbee.2013.08.016]
45. Vuoti AO, Tulppo MP, Ukkola OH, Junttila MJ, Huikuri HV, Kiviniemi AM, et al. Prognostic value of heart rate variability in patients with coronary artery disease in the current treatment era. PLoS One 2021;16(7):e0254107 [FREE Full text] [doi: 10.1371/journal.pone.0254107] [Medline: 34214132]
46. Devi R, Tyagi HK, Kumar D. Heart rate variability analysis for early stage prediction of sudden cardiac death. World Academy of Science, Engineering and Technology. 2007. URL: https://publications.waset.org/10004326/heart-rate-variability-analysis-for-early-stage-prediction-of-sudden-cardiac-death [accessed 2021-05-14]
47. Park SJ, Cho K, Kwon O, Park H, Lee Y, Shim WH, et al. Development and validation of a deep-learning-based pediatric early warning system: a single-center study. Biomed J 2021 Jan:In Press. [doi: 10.1016/j.bj.2021.01.003]
48. Jang D, Kim J, Jo YH, Lee JH, Hwang JE, Park SM, et al. Developing neural network models for early detection of cardiac arrest in emergency department. Am J Emerg Med 2020 Jan;38(1):43-49. [doi: 10.1016/j.ajem.2019.04.006] [Medline: 30982559]
49. Tylman W, Waszyrowski T, Napieralski A, Kamiński M, Trafidło T, Kulesza Z, et al. Real-time prediction of acute cardiovascular events using hardware-implemented Bayesian networks. Comput Biol Med 2016 Feb 01;69:245-253. [doi: 10.1016/j.compbiomed.2015.08.015] [Medline: 26456181]
50. Somanchi S, Adhikari S, Lin A, Eneva E, Ghani R. Early prediction of cardiac arrest (Code Blue) using electronic medical records. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015 Presented at: KDD '15: The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; August 10 - 13, 2015; Sydney NSW Australia p. 2119-2126. [doi: 10.1145/2783258.2788588]
51. Kennedy CE, Aoki N, Mariscalco M, Turley JP. Using time series analysis to predict cardiac arrest in a PICU. Pediatr Crit Care Med 2015;16(9):332-339. [doi: 10.1097/pcc.0000000000000560]

52. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the ward. Crit Care Med 2016 Feb;44(2):368-374 [FREE Full text] [doi: 10.1097/CCM.0000000000001571] [Medline: 26771782]
53. Badriyah T, Briggs JS, Meredith P, Jarvis SW, Schmidt PE, Featherstone PI, et al. Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). Resuscitation 2014 Mar;85(3):418-423. [doi: 10.1016/j.resuscitation.2013.12.011] [Medline: 24361673]
54. Ho JC, Park Y. Learning from different perspectives: robust cardiac arrest prediction via temporal transfer learning. In: Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2017 Presented at: 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC); July 11-15, 2017; Jeju, Korea (South). [doi: 10.1109/embc.2017.8037162]
55. Hong S, Lee S, Lee J, Cha WC, Kim K. Prediction of cardiac arrest in the emergency department based on machine learning and sequential characteristics: model development and retrospective clinical validation study. JMIR Med Inform 2020 Aug 04;8(8):e15932 [FREE Full text] [doi: 10.2196/15932] [Medline: 32749227]
56. Cho K, Kwon O, Kwon J, Lee Y, Park H, Jeon K, et al. Detecting patient deterioration using artificial intelligence in a rapid response system. Crit Care Med 2020;48(4):285-289. [doi: 10.1097/ccm.0000000000004236]
57. Liu JH, Chang HK, Wu CT, Lim WS, Wang HC, Jang JS. Machine learning based early detection system of cardiac arrest. In: Proceedings of the International Conference on Technologies and Applications of Artificial Intelligence (TAAI). 2019 Presented at: International Conference on Technologies and Applications of Artificial Intelligence (TAAI); Nov. 21-23, 2019; Kaohsiung, Taiwan. [doi: 10.1109/taai48200.2019.8959922]
58. Javan SL, Sepehri MM, Javan ML, Khatibi T. An intelligent warning model for early prediction of cardiac arrest in sepsis patients. Comput Methods Programs Biomed 2019 Sep;178:47-58. [doi: 10.1016/j.cmpb.2019.06.010] [Medline: 31416562]
59. Majumder AJ, ElSaadany YA, Young R, Ucci DR. An energy efficient wearable smart IoT system to predict cardiac arrest. Adv Hum Comput Interact 2019 Feb 12;2019:1-21. [doi: 10.1155/2019/1507465]
60. Kim J, Chae M, Chang H, Kim Y, Park E. Predicting cardiac arrest and respiratory failure using feasible artificial intelligence with simple trajectories of patient data. J Clin Med 2019 Aug 29;8(9):1336 [FREE Full text] [doi: 10.3390/jcm8091336] [Medline: 31470543]
61. Akrivos E, Papaioannou V, Maglaveras N, Chouvarda I. Prediction of cardiac arrest in intensive care patients through machine learning. In: Maglaveras N, Chouvarda I, de Carvalho P, editors. Precision Medicine Powered by pHealth and Connected Health. Singapore: Springer; 2018:25-29.
62. Kwon J, Lee Y, Lee Y, Lee S, Park J. An algorithm based on deep learning for predicting in‐hospital cardiac arrest. J Am Heart Assoc 2018 Jul 03;7(13):e008678. [doi: 10.1161/jaha.118.008678]
63. Rohila A, Sharma A. Detection of sudden cardiac death by a comparative study of heart rate variability in normal and abnormal heart conditions. Biocybern Biomed Eng 2020 Jul 03;40(3):1140-1154. [doi: 10.1016/j.bbe.2020.06.003]
64. Raka A, Naik G, Chai R. Computational algorithms underlying the time-based detection of sudden cardiac arrest via electrocardiographic markers. Appl Sci 2017 Sep 16;7(9):954. [doi: 10.3390/app7090954]
65. Murugappan M, Murukesan L, Omar I, Khatun S, Murugappan S. Time domain features based sudden cardiac arrest prediction using machine learning algorithms. J Med Imaging Hlth Inform 2015 Nov 01;5(6):1267-1271. [doi: 10.1166/jmihi.2015.1525]
66. Ebrahimzadeh E, Pooyan M, Bijar A. A novel approach to predict sudden cardiac death (SCD) using nonlinear and time-frequency analyses from HRV signals. PLoS One 2014 Feb 4;9(2):e81896 [FREE Full text] [doi: 10.1371/journal.pone.0081896] [Medline: 24504331]
67. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Risk of mortality and cardiopulmonary arrest in critical patients presenting to the emergency department using machine learning and natural language processing. PLoS One 2020 Apr 2;15(4):e0230876 [FREE Full text] [doi: 10.1371/journal.pone.0230876] [Medline: 32240233]
68. Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Syst Appl X 2020 Sep;7:100033. [doi: 10.1016/j.eswax.2020.100033]
69. Sahoo S, Dash M, Behera S, Sabut S. Machine learning approach to detect cardiac arrhythmias in ECG signals: a survey. IRBM 2020 Aug;41(4):185-194. [doi: 10.1016/j.irbm.2019.12.001]
70. Lonsdale H, Jalali A, Ahumada L, Matava C. Machine learning and artificial intelligence in pediatric research: current state, future prospects, and examples in perioperative and critical care. J Pediatr 2020 Jun;221S:3-10. [doi: 10.1016/j.jpeds.2020.02.039] [Medline: 32482232]
71. 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]
72. Itchhaporia D. Artificial intelligence in cardiology. Trends Cardiovasc Med 2020 Nov 23:A. [doi: 10.1016/j.tcm.2020.11.007] [Medline: 33242635]

Abbreviations

AI: artificial intelligence
DL: deep learning
DT: decision tree
HRV: heart rate variability
KNN: k-nearest neighbors
LR: logistic regression
MeSH: Medical Subject Headings
ML: machine learning
OHCA: out-of-hospital cardiac arrest
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
SVM: support vector machine

Edited by C Lovis; submitted 28.05.21; peer-reviewed by J Walsh, P Sarajlic; comments to author 23.09.21; revised version received 07.10.21; accepted 10.10.21; published 17.12.21.

Please cite as:
Alamgir A, Mousa O, Shah Z
Artificial Intelligence in Predicting Cardiac Arrest: Scoping Review
JMIR Med Inform 2021;9(12):e30798
URL: https://medinform.jmir.org/2021/12/e30798
doi:10.2196/30798
PMID:34927595

Review

Artificial Intelligence–Based Framework for Analyzing Health Care Staff Security Practice: Mapping Review and Simulation Study

Prosper Kandabongee Yeng[1], MSc; Livinus Obiora Nweke[1], MSc; Bian Yang[1], PhD; Muhammad Ali Fauzi[1], MSc; Einar Arthur Snekkenes[1], PhD

Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Gjovik, Norway

Corresponding Author:
Prosper Kandabongee Yeng, MSc
Department of Information Security and Communication Technology
Norwegian University of Science and Technology
Teknologivegen 22
Gjovik, 2815
Norway
Phone: 47 61135400
Email: prosper.yeng@ntnu.no

Abstract

Background: Blocklisting malicious activities in health care is challenging in relation to access control in health care security practices due to the fear of preventing legitimate access for therapeutic reasons. Inadvertent prevention of legitimate access can contravene the availability trait of the confidentiality, integrity, and availability triad, and may result in worsening health conditions, leading to serious consequences, including deaths. Therefore, health care staff are often provided with a wide range of access such as a "breaking-the-glass" or "self-authorization" mechanism for emergency access. However, this broad access can undermine the confidentiality and integrity of sensitive health care data because breaking-the-glass can lead to vast unauthorized access, which could be problematic when determining illegitimate access in security practices.

Objective: A review was performed to pinpoint appropriate artificial intelligence (AI) methods and data sources that can be used for effective modeling and analysis of health care staff security practices. Based on knowledge obtained from the review, a framework was developed and implemented with simulated data to provide a comprehensive approach toward effective modeling and analyzing security practices of health care staff in real access logs.

Methods: The flow of our approach was a mapping review to provide AI methods, data sources and their attributes, along with other categories as input for framework development. To assess implementation of the framework, electronic health record (EHR) log data were simulated and analyzed, and the performance of various approaches in the framework was compared.

Results: Among the total 130 articles initially identified, 18 met the inclusion and exclusion criteria. A thorough assessment and analysis of the included articles revealed that K-nearest neighbor, Bayesian network, and decision tree (C4.5) algorithms were predominantly applied to EHR and network logs with varying input features of health care staff security practices. Based on the review results, a framework was developed and implemented with simulated logs. The decision tree obtained the best precision of 0.655, whereas the best recall was achieved by the support vector machine (SVM) algorithm at 0.977. However, the best F1-score was obtained by random forest at 0.775. In brief, three classifiers (random forest, decision tree, and SVM) in the two-class approach achieved the best precision of 0.998.

Conclusions: The security practices of health care staff can be effectively analyzed using a two-class approach to detect malicious and nonmalicious security practices. Based on our comparative study, the algorithms that can effectively be used in related studies include random forest, decision tree, and SVM. Deviations of security practices from required health care staff's security behavior in the big data context can be analyzed with real access logs to define appropriate incentives for improving conscious care security practice.

(JMIR Med Inform 2021;9(12):e19250) doi:10.2196/19250

KEYWORDS

artificial intelligence; machine learning; health care; security practice; framework; security; modeling; analysis

Introduction

Background

Unlike other sectors, the health care sector cannot afford to implement stricter control for accessing sensitive health care information for therapeutic purposes. Despite the recognized need to provide tighter security measures in controlling access, there is also the need to strike a balance for allowing legitimate access to health care data for therapeutic reasons [1,2]. In access control management in health care, access to personal health data and personal data filing systems for therapeutic purposes must be granted following a specific decision based on "the completed or planned implementation of measures for the medical treatment of the patient" [3]. Therefore, access must only be granted to those with official needs [3,4]. While providing restrictions against unauthorized access, there are some provisions for following the availability trait of the confidentiality, integrity, and availability (CIA) triad during emergency situations. These include the provision for self-authorization. Self-authorization, or "break-the-glass," is a "technical measure which has been established for health personnel to be able to gain access to personal health data and personal data as and when necessary" [1]. However, access through self-authorization must be verified for abuse, and clear misuse must be followed up as a data breach [3,5].

The challenge remains in detecting misuse over a broad range of access [1,2]. A broad range of access via self-authorization results in tones of variant data known as "big data" [6], making it complex to manually determine legitimate access. However, in light of the recent increase in data breaches within health care, it has become necessary to adopt state-of-the-art methods to determine anomalous access. In the Healthcare Security Practice Analysis, Modeling, and Incentivization (HSPAMI) project [7], data-driven and artificial intelligence (AI) approaches were identified and adopted to aid in modeling and analyzing health care staff's security practices in their access control logs [7]. AI is based on algorithms in computer science that can be used for analyzing complex data to draw meaningful patterns and relationships toward decision making [8]. The aim of this study was to understand anomaly practices in health care in the context of big data and AI, and to determine the security practice challenges often faced by health care workers while performing their duties. The results will provide knowledge to serve as a guide for finding better approaches to security practice in health care. However, there are different types of data sources and AI methods that can be used in this approach [7]. We therefore adopted a review methodology to first detail various types of dimensions, including the data sources and AI methods, which can be adopted in related studies.

According to Verizon, the health care sector globally experienced approximately 503 data breaches in 2018, which resulted in the compromise of up to 15 million records [4,9]. This figure was triple the number of data breaches recorded in 2017. In addition, the number of records compromised within the health care sector in 2019 far exceeded that recorded in 2018 [9]. Unfortunately, more than half of these data breaches were perpetuated by insiders [9]. The report opined that approximately 83% of the adversaries were motivated by financial gains, 3% were due to convenience, 3% were due to grudges, and 2% were a result of industrial espionage. The current situation implies that the number of data breaches within the health care sector has surpassed that of the financial sector and almost equals those of other public sectors.

This situation has raised concerns among relevant stakeholders, and many are wondering the reasons behind the spike in the number of data breaches within the health care sector. Some of these reasons can be easily deduced because health care data have economic value and as such represent a possible target for malicious actors [10,11]. Moreover, health care data have scientific and societal value that makes them very attractive for cyber criminals. In fact, Garrity et al [12] indicated that patient medical records are sold for approximately US $1000 on the dark web. Another reason for data breaches within health care is the lack of health care personnel. The few health care personnel are more interested in their core health care duties and have little time to handle health care information security issues. This situation provides cyber criminals with the opportunity to exploit health care systems.

Although there have been improvements in technical measures, such as firewalls, intrusion detection and prevention systems, antivirus software, and security governance configurations, the development of a "human firewall" has not been considered [13,14]. The "human firewall" refers to the information security conscious care behavior of insiders [15]. However, this concept has not received equal attention as devoted to technical measures, and thus cyber criminals seek to exploit it for easy access [16]. Health care insiders have access privileges that enable them to provide therapeutic care to patients; however, through errors or deliberate actions, they can compromise the CIA of health care data. It is also possible for an attacker to masquerade as an insider to compromise health care data through social engineering and other methods [17,18].

Access control mechanisms within the health care sector are usually designed with a degree of flexibility to facilitate efficient patient management [19]. Even though such design considerations are vital and can meet the availability attribute of the CIA, they make health care systems vulnerable. This is because flexibility can be abused by insiders [20]. In addition, an attacker who could obtain an insider's access privilege can exploit this flexibility to have broader access. A successful data breach could have many consequences such as denial of timely medical services, corrosion of trust between the patient and health care providers, breaches to an individual's privacy [21], and huge fines to health care providers by national and international regulatory bodies. The general objective of this study was to determine an effective way of modeling and analyzing health care logs. A review was first performed to retrieve appropriate data sources and their features in addition to identifying the AI methods that can best be used to determine irregularities in security practices among health care workers.

Prior Studies

The security practices of health care staff include how health care professionals respond to security controls and measures for achieving the CIA goals of health care organizations [2,4,5].

Health care professionals are required to conduct their work activities in a security-conscious manner to maintain the CIA of the health care environment [3]. For instance, borrowing access credentials could jeopardize the purpose of access control for authorized users and legitimate access. Additionally, the inability to understand social engineering scammers' behavior can lead to health care data breaches [7].

Various approaches can be adopted to observe, model, and analyze health care professionals' security practices. A perception and sociocultural context can be adopted by analyzing the security perception, and social, cultural, and sociodemographic characteristics of health care staff in the context of their required security practices [7,22]. In addition, an attack-defense simulation can be used to measure how health care staff understand social engineering–related tricks. Furthermore, a data-driven approach with AI methods could be adopted to understand the security behavior of each health care professional in the context of big data, since AI is most appropriate for analyzing complex data sets with high volume, variety, velocity, and veracity [8]. The findings can then help decision makers to introduce appropriate incentive methods and solve issues that hinder sound information security practice toward enhancing conscious care behavior.

Advances in computational and data science, along with engineering innovations in medical devices, have prompted the need for the application of AI in the health care sector [23-25]. This has the potential to improve health care delivery and revolutionize the health care industry. AI can be referred to as the use of complex algorithms and software to imitate human cognitive functions [24-26]. AI involves the application of computer algorithms in the process of extracting meaning from complicated data and making intelligent decisions without direct human input [24,25]. AI is increasingly impacting every aspect of our lives, and the health care sector is no exception. In recent years, the health care sector experienced massive AI deployments in the bid to improve overall health care delivery. We here rely on the classification of the application of AI in health care described by Wahl et al [27] to briefly discuss the deployment of AI in health care.

According to Wahl et al [27], the deployment of AI in the health care sector has been classified to include expert systems, machine learning, natural language processing, automated planning and scheduling, and image and signal processing [27]. Expert systems are AI programs that have been trained with real cases to execute complicated tasks [28]. Machine learning employs algorithms to identify patterns in data and learn from them, and its applications can be grouped into three categories: supervised learning, unsupervised learning, and reinforcement learning [25,27]. Natural language processing facilitates the use of AI to determine the meaning of a text by using algorithms to identify keywords and phrases in natural language. Automated planning and scheduling is an emerging field in the use of AI in health care that is concerned with the organization and prioritization of the necessary activities to obtain the desired aim [27]. Image and signal processing involves the use of AI to train information extracted from a physical occurrence (images and signals) [27].

The common characteristic of all these applications is the utilization of massive data that are being generated in the health care sector to make better informed decisions. For instance, the collection of data generated by health care staff has been used for disease surveillance, decision support systems, detecting fraud, and enhancing privacy and security [29]. In fact, the code of conduct for the Norwegian health care sector requires the appropriate storage and protection of access logs of health care information systems for security reasons [3]. Health care staff's access to the network or electronic health records (EHR) leaves traces of their activities, which can be logged and reconstructed to form their unique profiles [3,4]. Therefore, appropriate AI methods can be used to mine such logs to determine the unique security practices of health care staff. Such findings can support management in adapting suitable incentivization methods toward improving security-conscious care behavior in health care. Therefore, the aim of this study was to explore the appropriate AI methods and data sources that can be used to observe, model, and analyze the security practices of health care staff.

HSPAMI is an ongoing research project with one aspect involving the modeling and analysis of data with AI methods to determine the security practices of health care staff toward improving their security-conscious care behavior. In analyzing health care–related data, there is a need to consider details of the methods and data sources in view of the unique and critical nature of the sector. In a related study, Walker-Roberts et al [30] performed a systematic review of "the availability and efficacy of countermeasures to internal threats in health care critical infrastructure." Among various teams, few machine learning methods were identified to be used for intrusion detection and prevention. The methods that were identified are Petri net, fuzzy logic, k-nearest neighbor (KNN), decision tree (RADISH system) [30-32], and inductive machine learning methods [30,31,33]. In a similar way, Islam et al [34] performed a systematic review on data mining for health care analytics. Categories such as health care subareas; data mining techniques; and the types of analytics, data, and data sources were considered in the study. Most of the data analysis was focused on clinical and administrative decision-making. The data sources were mostly human-generated from EHRs. Gheyas et al [35] also explored related methods in their systematic review and meta-analysis [35].

Even though the studies of Walker-Roberts et al [30] and Islam et al [34] were in the health care context, details of the algorithms and data sources were not considered. For instance, the features of the data sources and algorithm performance methods were not deeply assessed in their studies. Additionally, these studies were general and not specific to health care [35,36], and therefore the unique challenges within the health care environment were not considered. To this end, this study explored AI methods and data sources in health care that can be efficiently used for modeling and analyzing health care professionals' behavior. The terms "health care professionals" and "health care staff" are used interchangeably in this paper, which include, but are not limited to, nurses, physicians, laboratory staff, and pharmacies who access patient records for therapeutic reasons.

Scope, Problem Specification, and Contribution

Following the recent increase in data breaches in health care, our research group is working on the HSPAMI project, which was initiated to measure the information security practice level of health care staff [7,22]. The results will help provide better approaches for incorporating conscious care behavior among health care staff. The HSPAMI project has already identified various approaches to include psychosociocultural context attack and defense simulations in a social engineering context along with data-driven AI approaches [7].

The main goal is to demonstrate how health care security practices can be analyzed to determine anomalous and malicious activities in the context of data-driven and AI approaches. Therefore, the specific objectives of this study were to identify, assess, and analyze the state-of-the-art data-driven attributes and AI methods along with their design strategies and challenges. A framework for analyzing health care security practice in the context of data-driven and AI methods was also developed and evaluated. The broad goal was to enable analysis of real logs of health care professionals' security practices in the context of big data and human-generated data logs. Therefore, the psychosociocultural context and attack-defense simulations are beyond the scope of this paper.

Some details of data sources and AI methods that can be used in this study were not provided in previous related work [30-34], which raised several questions for our research: Among the various data sources that are generated by health care staff, which is the most appropriate to be used in analyzing the security practice? Which AI methods have been pinpointed to be suitable for use in modeling and analyzing health care security practice? What evaluation techniques are most appropriate in this context, and how were these methods adjusted to curtail biases amid various access points, such as self-authorization during emergency care scenarios and the busy schedules of health care staff? To answer these questions, we first performed a mapping review [37] toward identifying, modeling, and analyzing health care staff–generated access logs and AI methods to enhance security practice. This work represents an extended version of our previous work, with the additions being a design and framework evaluation.

Methods

Literature Review

Various types of systematic studies exist [38-41], including a systematic mapping study, scoping review, and systematic literature review. Systematic mapping studies review topics with a broader scope by categorizing the identified research articles into specific areas of interest. Systematic mapping studies have general research questions with the objective to determine research trends or the state-of-the-art studies. By contrast, the objective of a systematic literature review is to accumulate data and therefore has a more specific research focus. To this end, a systematic mapping study was adopted in this work [38,39]. Based on the results, we developed a framework that was evaluated with simulated log data.

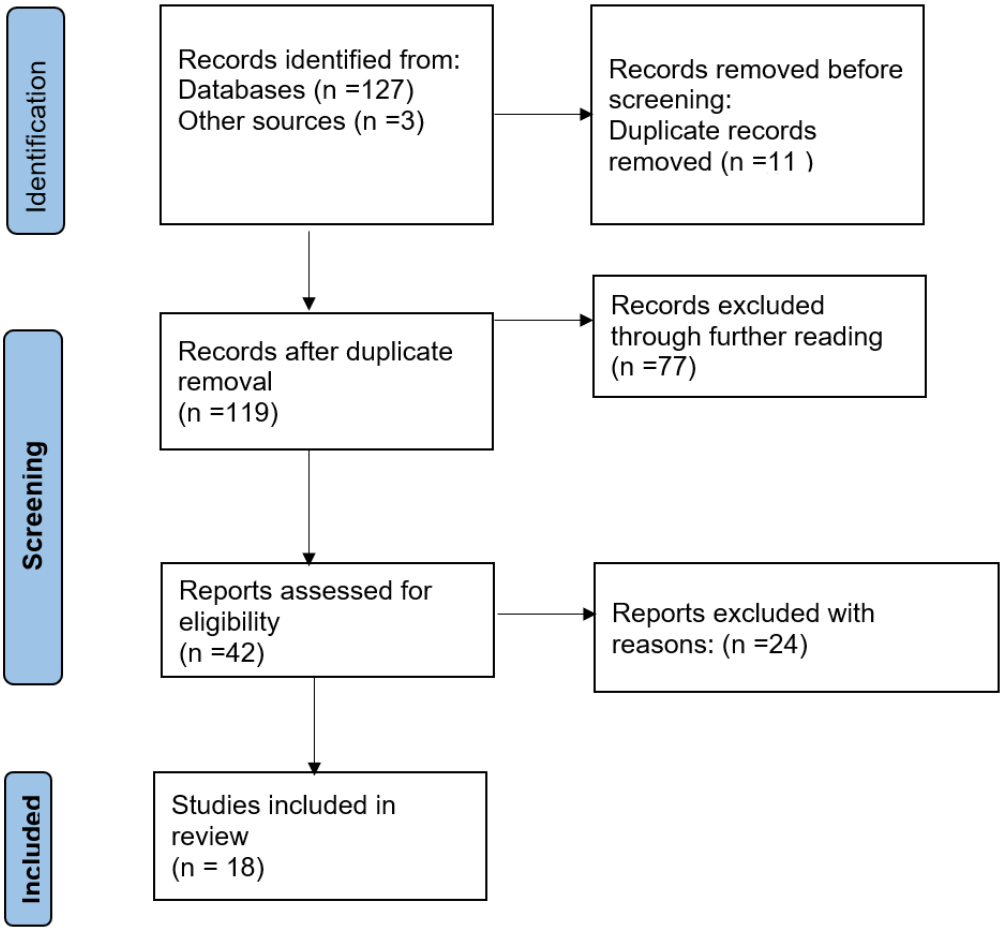
Although we did not restrict the article search to a specific time frame, we performed the literature search between June 2019 and December 2019 with the Google Scholar, Science Direct, Elsevier, IEEE Explore, ACM Digital, Scopus, Web of Science, and PubMed databases. Different keywords were used, including "healthcare," "staff," "employee," "information security," "behavior," "practice," "threat," "anomaly detection," "intrusion detection," "artificial intelligence," and "machine learning." To ensure a high-quality searching approach, the keywords were combined using the Boolean functions "AND," "OR," and "NOT." For instance, the following search string was generated in PubMed:

((Intrusion[All Fields] AND Detection[All Fields]) OR (Anomaly[All Fields] AND Detection[All Fields])) AND ("health"[MeSH Terms] OR "health"[All Fields]) AND (("artificial intelligence"[MeSH Terms] OR ("artificial"[All Fields] AND "intelligence"[All Fields]) OR "artificial intelligence"[All Fields]) OR ("machine learning"[MeSH Terms] OR ("machine"[All Fields] AND "learning"[All Fields]) OR "machine learning"[All Fields])) AND ("information"[All Fields] AND Security[All Fields]) AND (("behavior"[All Fields] OR "behavior"[MeSH Terms] OR "behavior"[All Fields]) OR "practice"[All Fields]).

Peer-reviewed articles were considered. The inclusion and exclusion criteria were developed based on the objective of the study and through rigorous discussions among the authors.

Basic selection was performed by initially skimming through the titles, abstracts, and keywords to retrieve records that were in line with the inclusion and exclusion criteria. Duplicates were filtered out, and articles that seemed relevant, based on the inclusion and exclusion criteria, were fully read and evaluated. Each of the authors independently read and assessed all of the selected articles and judged either to be included or excluded. Using the inclusion and exclusion criteria as a guideline, discrepancies were discussed and resolved among the authors. Other appropriate articles were also retrieved using the reference list of accepted literature. Figure 1 shows the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) [42] flowchart of article screening and selection.

Figure 1. Flowchart of the systematic review process.



Inclusion and Exclusion Criteria

For an article to be included in the review, it had to be related to anomaly detection or intrusion detection in health care using AI methods with health care professional–generated access log data or patterns. Any other article outside the above scope (such as articles related to medical cyber-physical devices, body area networks, and similar), along with articles published in languages other than English, were excluded.

Data Collection and Categorization

The data collection and categorization methods were developed based on the study objective, and thorough literature reviews and discussions among the authors. The categories were defined exclusively to assess, analyze, and evaluate the study objectives, which are summarized in Table 1.

Table 1. Data categories and their exclusive definitions.

Category	Definition	Examples
Type of AI[a] method	Explicit machine learning methods	Support vector machine, Bayesian network
Type of input	Features used by the algorithm	Access location, time, failed login attempts
Input sources	Type of access log data used in the study	Browser history, network logs, host-based activity logs, EHR[b] logs
Data format, type, size, and data source	File formats	XML, comma separated value (CSV)
Input preprocessing	Defines how the data were preprocessed and how missing and corrupted input data were handled	Structured vs unstructured
Security failures	Context in which the algorithm was implemented	Intrusion or anomaly detection
Ground truth	Type of training set used in training the model	Login and logout time, average number of patient records accessed
Privacy approach	Defines the privacy method used to safeguard the privacy rights of individuals who contributed to the data source	Message Digest 5 (MD5), Secure Hash Algorithm (SHA)-3
Performance metrics or evaluation criteria	Measures used to assess the accuracy of the study	Specificity, sensitivity, receiver operating characteristic curve
Nature of data sources	Specifies whether the data used were synthetic or real data	Real data, simulated data

[a]AI: artificial intelligence.

[b]EHR: electronic health record.

Literature Evaluation and Analysis

The selected articles were assessed, analyzed, and evaluated based on the categories defined in Table 1. The analysis was performed on each of the categories (eg, type of AI method, type of input, input source, preprocessing, learning techniques, performance methods) to evaluate the state-of-the-art approaches. Percentages of the attributes of the categories were calculated based on the total number of counts (n) of each type of attribute. Some studies used multiple categories; therefore, the number of counts of these categories exceeded the total number of articles of these systems presented in the study.

Results

Review Findings

Articles Retrieved

After searching the various online databases, a total of 130 records were initially identified following the guidelines of the inclusion and exclusion criteria in the reading of titles, abstracts, and keywords. A further assessment of these articles through skimming of the objective, method, and conclusion sections led to an exclusion of 77 articles that did not meet the defined inclusion criteria. After removing duplicates, 42 articles were fully read and judged. After full-text reading, a total of 18 articles were included in the study and analysis (Figure 1).

Algorithms

The main findings of the reviewed articles and their related categorizations such as algorithms, features, and data sources are shown in Figure 2. The algorithms, features, data sources, and application domains were the most frequent categorizations in the review; the study column presents the sources of each of these categories.

The algorithms that were most commonly used for analyzing security practice in the review are shown in Table 2. The KNN method was the most frequently used, followed by the Bayesian network and C4.5 decision tree.

Figure 2. Algorithms, features, related data sources, and application domain. KNN: k-nearest neighbor; SVM: support vector machine; EHR: electronic health record.

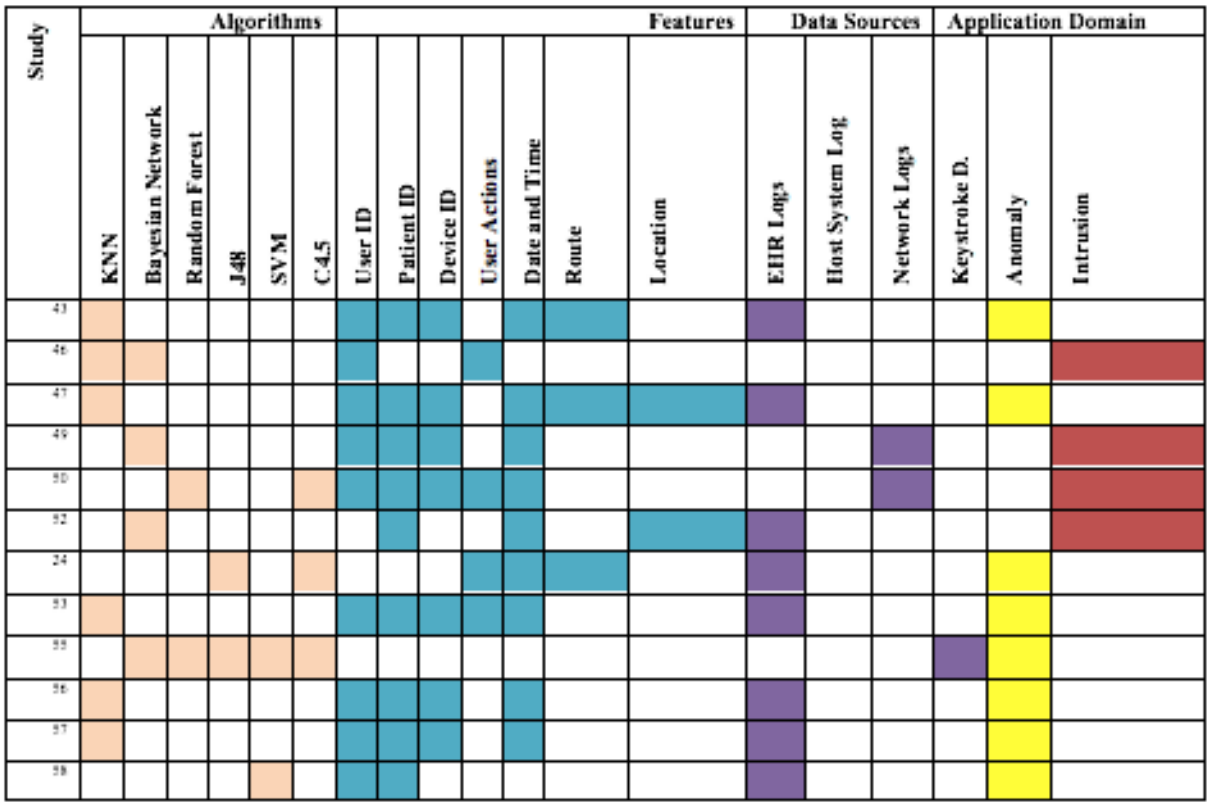


Table 2. Algorithms and their respective proportions among the articles included in the review (N=30).

Algorithm	Studies, n (%)	References
K-nearest neighbor	5 (17)	[43-47]
Bayesian network	4 (13)	[43,44,48,49]
Decision tree (C4.5)	3 (10)	[24,49,50]
Random forest	2 (7)	[49,50]
J48	2 (7)	[24,49]
Support vector machine	1 (3)	[49,51]
Spectral projection model	1 (3)	[47]
Principal component analysis	1 (3)	[47]
K-means	1 (3)	[52]
Ensemble averaging and a human-in-the-loop model	1 (3)	[53]
Partitioning around Medoids with k estimation (PAMK)	1 (3)	[50]
Distance-based model	1 (3)	[54]
White-box anomaly detection system	1 (3)	[55]
C5.0	1 (3)	[50]
Hidden Markov model	1 (3)	[54]
Graph-based	1 (3)	[56]
Logistic regression	1 (3)	[51]
Linear regression	1 (3)	[51]
Fuzzy cognitive maps	1 (3)	[57]

Features

Table 3 shows the unique features identified in the review and their respective counts and proportions. The features that were the most frequently used included user ID, date and time attribute, patient ID, and device identification.

Table 3. Features used in the reviewed articles (N=65).

Feature	Count, n (%)
User identification	13 (20.0)
Patient identification	11 (16.9)
Device identification	9 (13.8)
Access control	5 (7.7)
Date and time	11 (16.69)
Location	4 (6.2)
Service/route	5 (7.7)
Actions (delete, update, insert, copy, view)	3 (4.6)
Roles	3 (4.6)
Reasons	1 (1.5)

Data Sources

The majority of the data sources were EHR logs (11/18, 61%), followed by host-based logs (2/18, 11%), network logs (4/18, 22%), and keystroke activities (1/18, 5%).

Performance Methods

Table 4 shows the various types of performance methods that were identified with their respective counts and proportions; recall and receiver operating characteristic curve were the most common metrics applied, whereas F-score and root mean square error were the least commonly applied.

Table 4. Performance methods used in the reviewed studies (N=25).

Performance methods	Studies, n (%)
Receiver operating characteristic (ROC) curve	5 (20)
Area under ROC curve	3 (12)
Recall (sensitivity)	5 (20)
Precision	4 (16)
Accuracy	2 (8)
True negative rate (specificity)	3 (12)
F-score	2 (8)
Root mean square error	1 (4)

Security Failures

The studies in the review were mostly applied for anomaly detection (12/18, 67%) and malicious intrusion detection (6/18, 33%).

File Format

Among the 4 articles that reported the file format, 2 (50%) used comma separated values [43,52] and the other 2 (50%) used the SQL file format [55,58].

Ground Truth

Eight of the 18 articles included in the review reported the ground truth, which was established with similarity measures (3/8, 38%), observed practices (3/8, 38%), and historical data of staff practices (2/8, 25%).

Privacy-Preserving Data Mining Approach

Privacy-preserving methods adopted in the included studies were tokenization [43], deidentification [45], and removal of medical information [24].

Nature of Data Source

The majority of studies (15/18, 83%) used real data for analysis, with the remaining (3/18, 17%) using synthetic data.

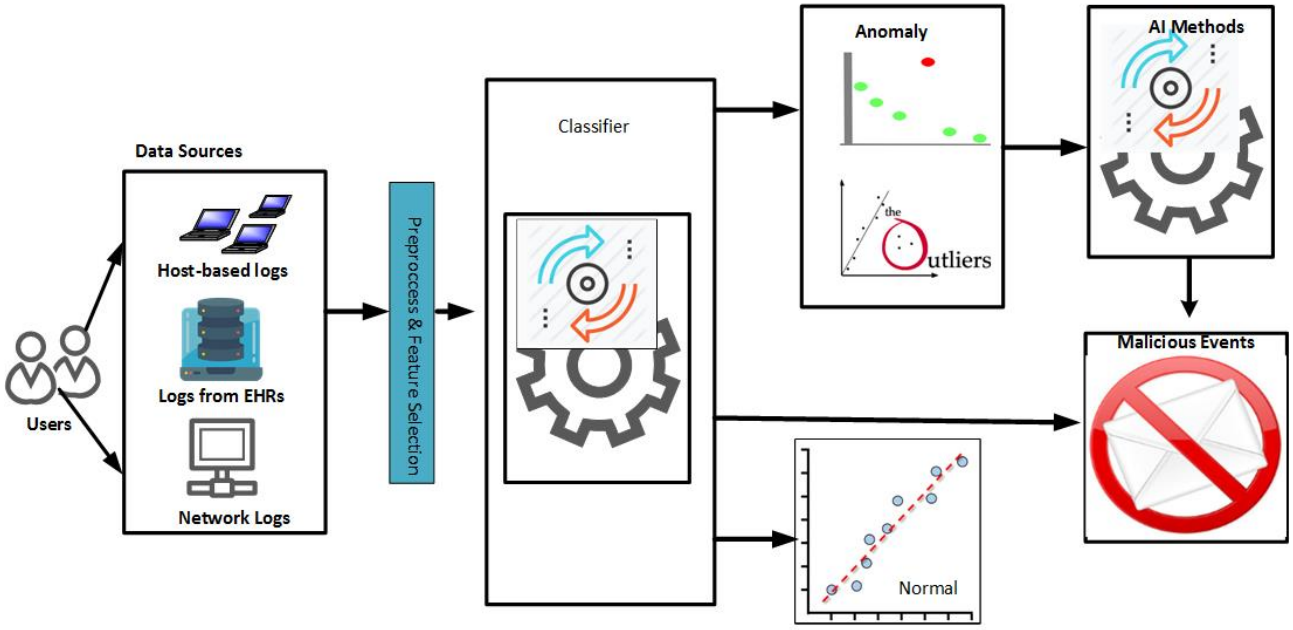
Framework for Analyzing Health Care Staff Security Practices

Based on the review, a conceptual framework was depicted on how data-driven and AI methods should be adopted to analyze logs of EHRs in security practice (see Figure 3). Our review indicated that a security practice analysis typically reveals the anomaly or malicious intrusion pattern of health care staff. Our model therefore has various dimensions such as data sources,

preprocessing, feature extraction, the application of AI methods, and possible classes, as shown in Figure 3.

The data sources include the network, EHR, or workstation logs. These logs are generated based on health care staff activities in accessing resources such as patients, printers, medical devices, and physical security systems. The logs go through the preprocessing phase [25], such as cleaning and feature selection. The essential features are then selected with appropriate methods, including filter methods, wrapper methods, or the combined filter and wrapper approach. Having obtained the appropriate features, a machine learning method can then be created, trained, and used to detect patterns of unusual security practices. The various classes that can be deduced in this framework include normal, abnormal, significantly nonmalicious anomaly, and malicious classes. The normal class includes features that follow the flow of each established access process without access aberration. The malicious class consists of features that violate established access flow and may also include excess access, which exceeds the usual trend of users. An example includes a doctor who accesses patient records more than the average daily access, and when the access was not for therapeutic measures. The anomaly nonmalicious class includes accesses that violate the established access flow or that exceed the average daily access of the health care staff; however, in this case, the accesses were for therapeutic purposes. From the framework, three access detection methods were identified for comparison.

Figure 3. Conceptual framework for analyzing the security practices of health care staff. AI: artificial intelligence; EHR: electronic health record.



Comparative Analysis of the Framework

The following three access detection methods were compared: (1) two-stage classification, (2) three-class classification, and (3) two-class classification. In the two-stage classification approach, the log data are classified as normal and anomaly. The data determined in the anomaly class from the first stage are further classified into two classes: malicious and nonmalicious (Figure 4). In the three-class approach, the log data are classified into normal, nonmalicious anomaly, and malicious, as shown in Figure 5. In the two-class approach, the normal and nonmalicious anomaly data are considered as a single "nonmalicious" category. The log data are then classified into nonmalicious and malicious classes, as shown in Figure 6.

These three approaches were then compared with nine machine learning methods: multinomial naive Bayes (NB), Bernoulli NB, Gaussian NB, KNN, neural network (NN), logistic regression (LR), random forest (RF), decision tree (DT), and support vector machine (SVM).

Figure 4. Flowchart of two-stage detection.

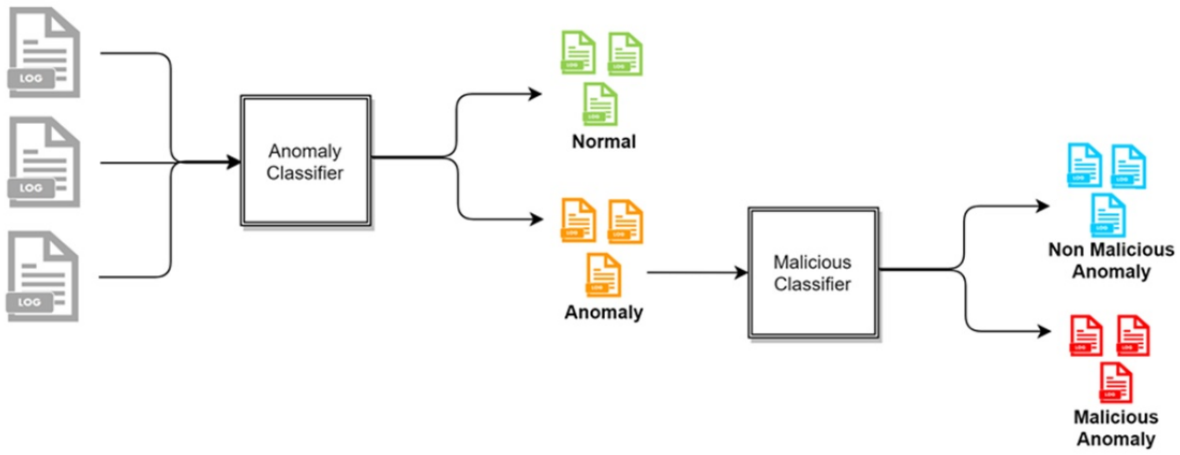


Figure 5. Two-class classification.

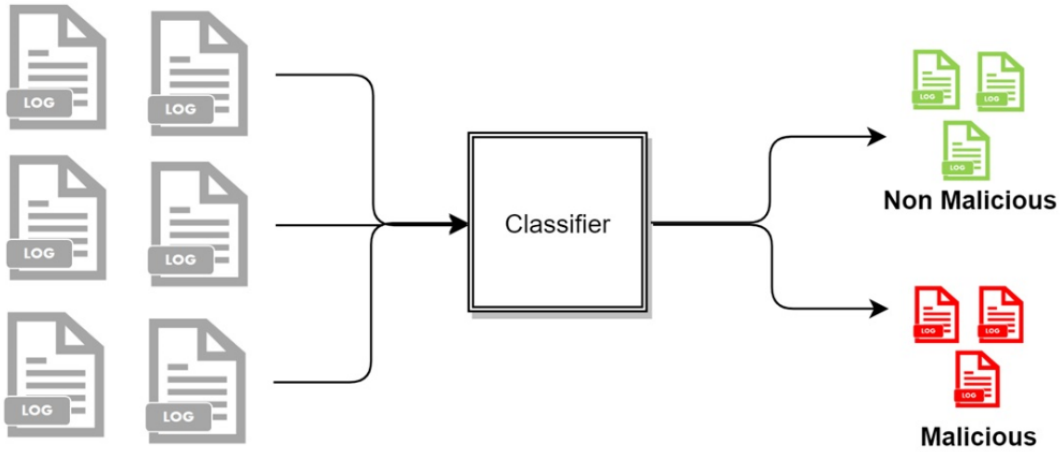
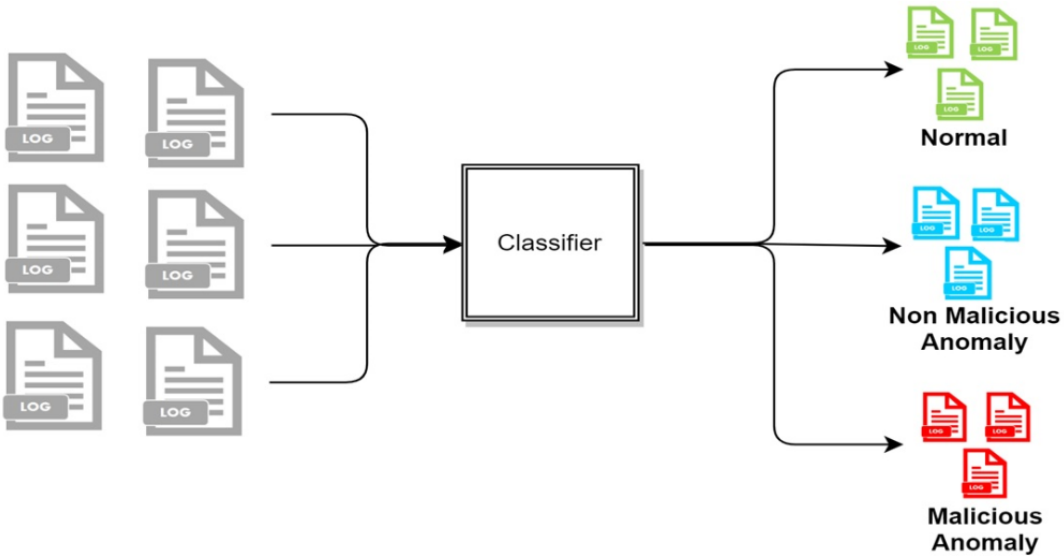


Figure 6. Three-class classification.



Simulation of EHR Logs of Health Care Staff Security Practice

The conceptual framework (Figure 3) provided direction and guidelines for effective modeling and analysis of health care staff security practices. We hence simulated 1-year access log data of a typical hospital information system from January 1, 2019, to December 31, 2019. Inpatient workflow, outpatient workflow, and emergency care patient workflow were modeled and used in the simulation of the logs as shown in Figure 7, Figure 8, and Figure 9, respectively. Five main modules were included in the simulation of the hospital information system:

Report, Finance, Patient Management, Laboratory Management, and Pharmacy Management. In the data simulation setting, we used 19 departments and 12 roles with a total of 53 employees. The departments were information technology (3 roles), finance (1 finance officer, 3 finance support staff), administration (1 head of administration, 2 support staff), pharmacy (3 roles), and medical laboratory (5 roles). Outpatient departments included ear-nose-throat (1 doctor, 2 nurses), dentistry (1 dentist, 2 nurses), pediatric unit (1 doctor), orthopedics (1 doctor, 2 nurses), neurology (1 doctor, 2 nurses), gynecology (1 doctor, 2 nurses), endocrinology (1 doctor, 2 nurses), rheumatology (1 doctor, 2 nurses), and cancer (1 doctor, 2 nurses). The inpatient departments included patient wards and the emergency department (2 doctors, 7 nurses).

Two types of shifts were used: a regular shift and three 8-hour shifts. The regular shift is Monday to Friday from 8 AM to 4 PM, whereas the three 8-hour shifts included the following three shifts every day of the week: (1) shift 1, 6 AM to 2 PM; (2) shift 2, 2 PM to 10 PM; and (3) shift 3, 10 PM to 6 AM (next day). The numbers of roles and employees in a regular shift and in the three 8-hour shifts are shown in Table 5.

Figure 7. Inpatient workflow.

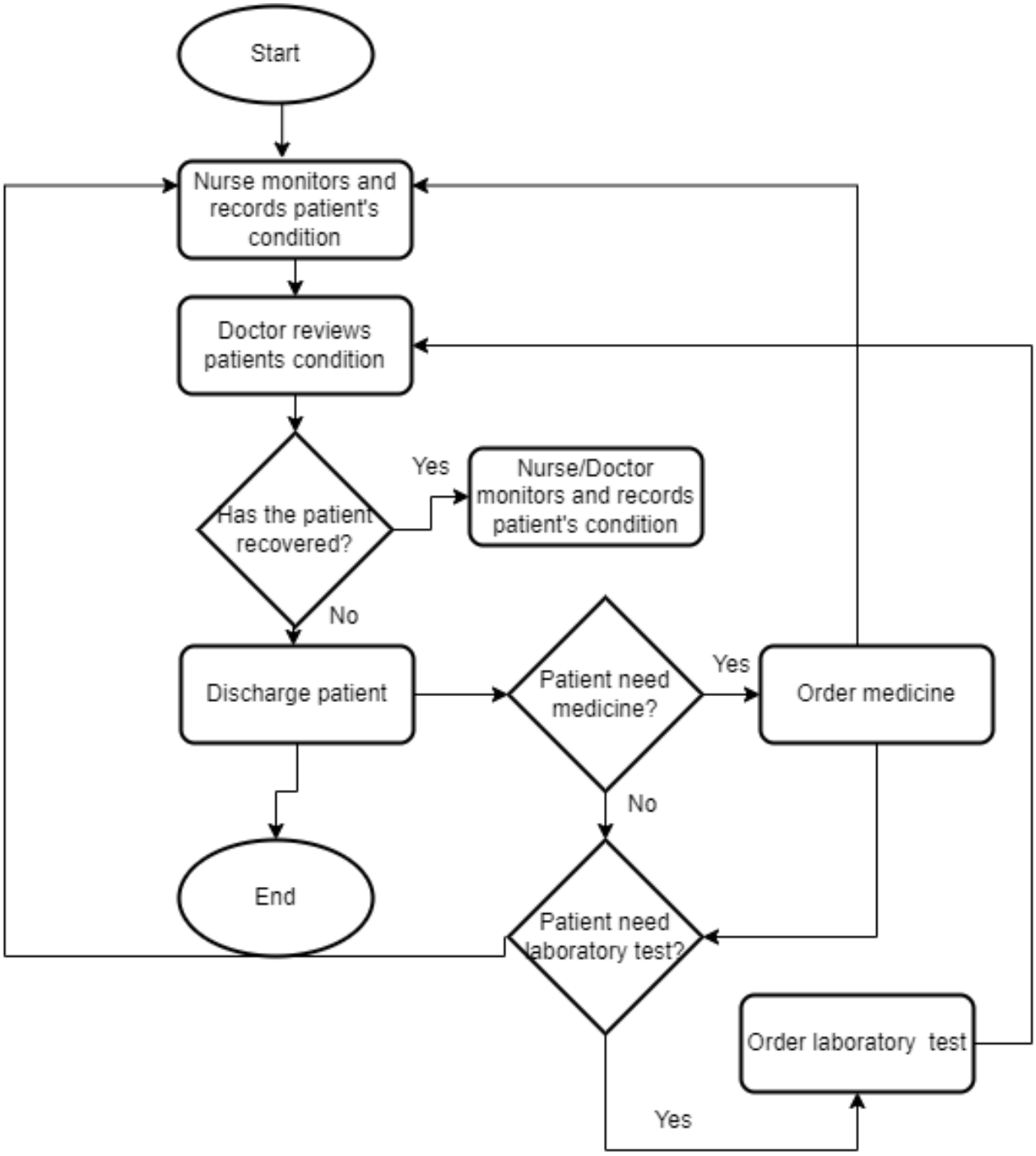


Figure 8. Emergency workflow.

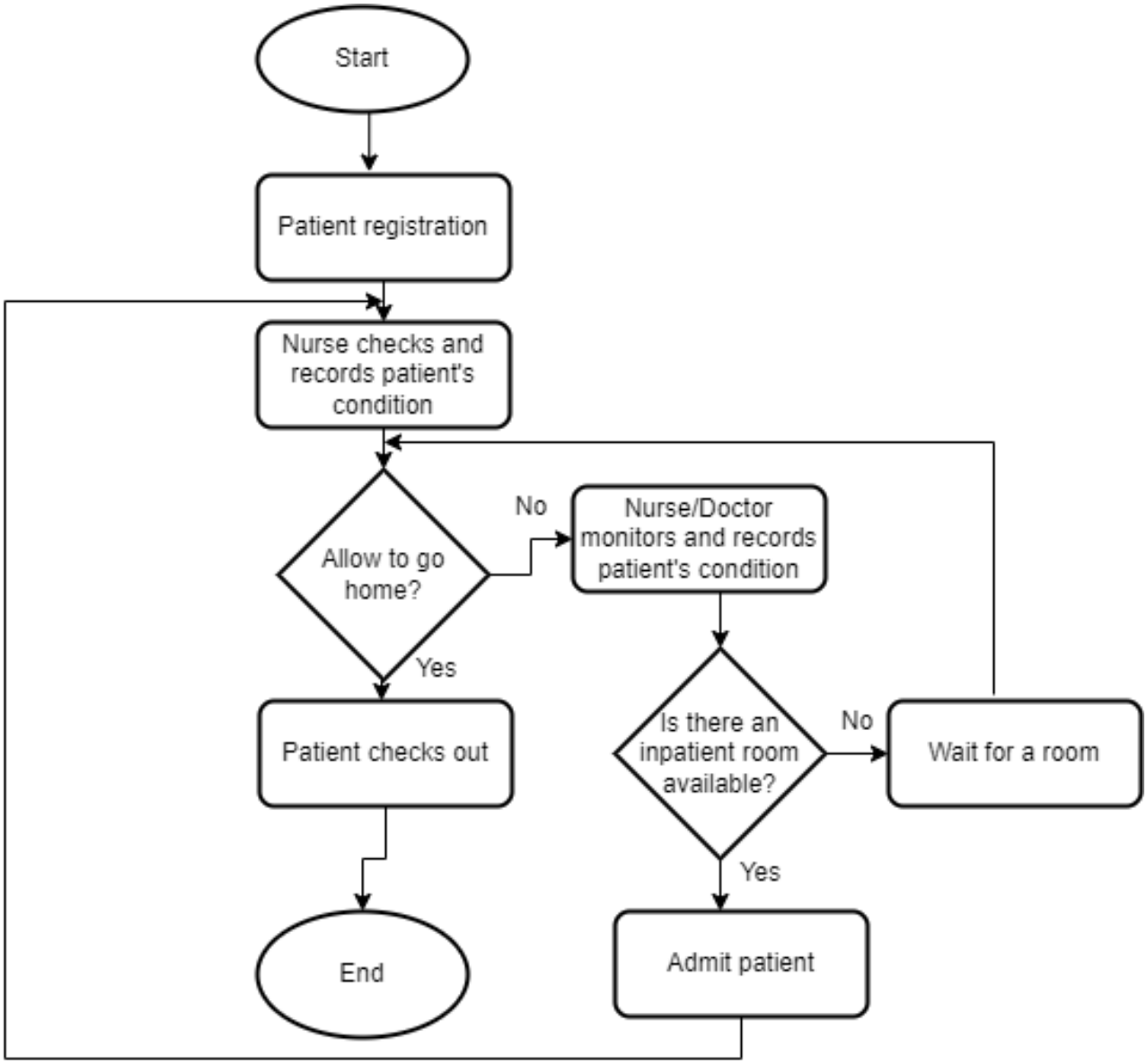


Figure 9. Outpatient care workflow.

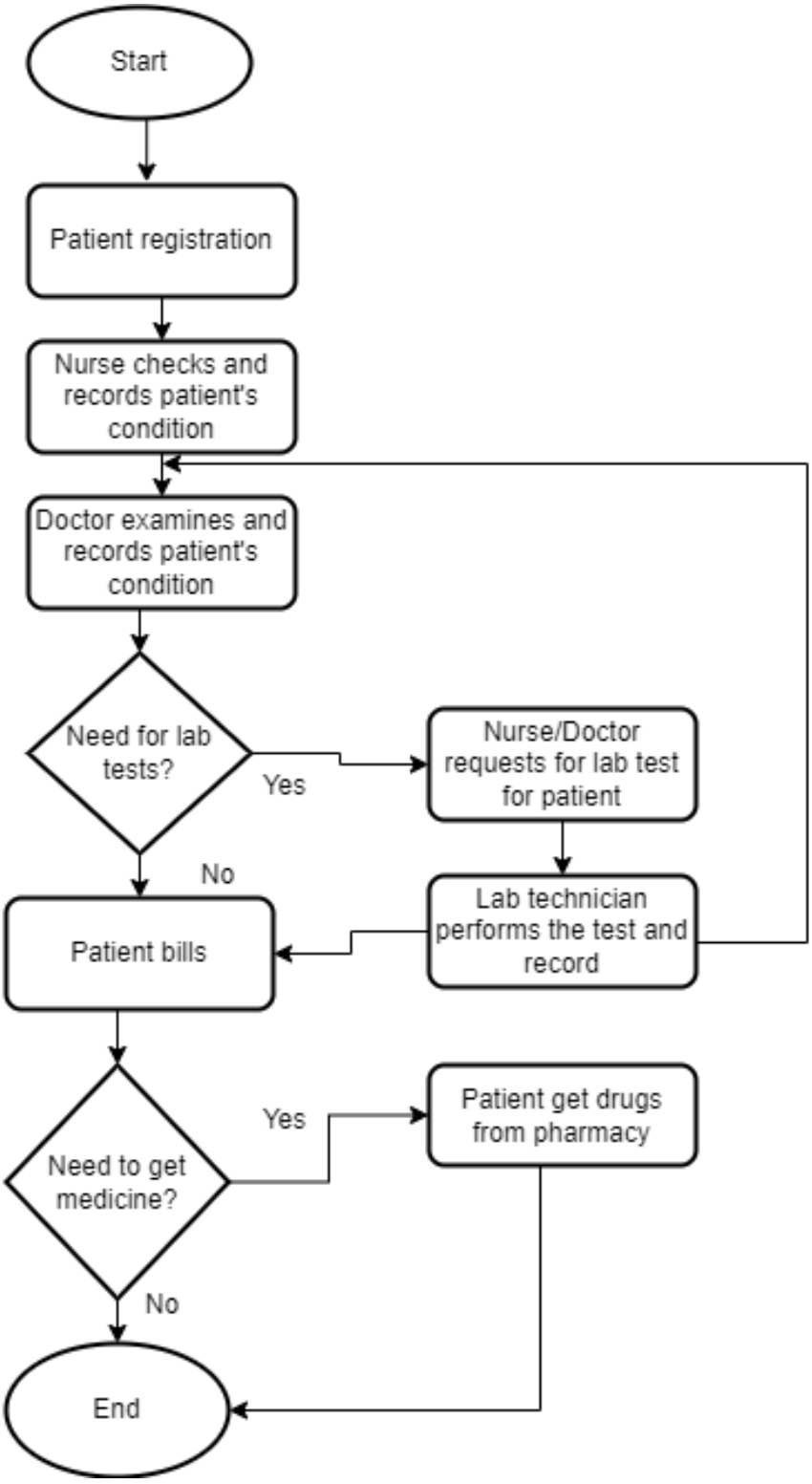


Table 5. Simulated departments, roles, and staff in a typical hospital.

Department	Roles (number of employees)
Information technology	Head (1), technical support (2)
Finance	Head (1), finance officer (4)
Administration	Head (1), administrative assistants (2)
Laboratory	Head (1), laboratory assistants (5)
Pharmacy	Head (1), pharmacy assistant (2)
Outpatient	
Ear-nose-throat	Doctor (1), nurse (2)
Optometry	Doctor (1), nurse (2)
Dentistry	Doctor (1), nurse (2)
Pediatrics	Doctor (1), nurse (2)
Orthopedics	Doctor (1), nurse (2)
Neurology	Doctor (1), nurse (2)
Gynecology	Doctor (1), nurse (2)
Endocrinology	Doctor (1), nurse (2)
Rheumatology	Doctor (1), nurse (2)
Cancer	Doctor (1), nurse (2)
Inpatient	
Ward 1	Doctor (1), nurse (2)
Ward 2	Doctor (1), nurse (2)
Ward 3	Doctor (1), nurse (2)
Three 8-hour shift	
Emergency	Doctor (2), nurse (2)
Ward 1	Nurse (2)
Ward 2	Nurse (2)
Ward 3	Nurse (2)

Based on the flows (see Figure 6 for an example), we simulated the data and recorded the logs. The logs are considered to be normal data (nonanomaly). We also simulated some abnormal data. The abnormal data were divided into two categories: nonmalicious and malicious. Nonmalicious abnormal data were generated by simulating the “break-the-glass” scenario (eg, access by a doctor from another department due to an emergency) [2], whereas malicious abnormal data were generated by simulating attackers that are assumed to have compromised some users’ credentials and used them to access patient records (eg, identity theft). In the latter category, the attacker will access more data than legitimate users and often not follow the flows. From this data simulation, 281,886 logs were created with 273,094 normal access, 7647 nonmalicious abnormal access, and 1145 malicious access scenarios. There are 21 fields recorded in this data simulation, as displayed in Table 6.

Table 6. Field attributes of simulated access logs of electronic health records.

Attribute	Description
startAccessTime	The time the employee starts to access the patient record: format=day/month/year, hours:minutes:seconds
endAccessTime	The time the employee ends the patient record access: format=day/month/year, hours:minutes:seconds
employeeID	The identification number of the employee who accesses the patient record (eg, record4roleID)
roleID	The role of the employee who accesses the patient record
patientID	The identification number of the patient whose record is being accessed by the employee
activityID	The identification number of the activity (1: Create, 2: Read, 3: Update, 4: Delete)
employeeDepartmentID	The department of the employee who accesses the patient record
employeeorganizationID	The organization of the employee who accesses the patient record
osID	The operating system of the computer used by the employee to access the patient record
deviceID	The identification number of the computer used by the employee to access the patient record
browserID	The browser used by the employee to access the patient record
ipAddress	The IP address of the computer used by the employee to access the patient record
ReasonID	The reason for the employee accessing the patient record (optional)
shiftID	The identification of the shift the employee belongs to on the day of accessing the patient record
shiftStartDate	The start time of the shift the employee belongs to on the day of accessing the patient record
shiftEndDateTime	The end time of the shift the employee belongs to on the day of accessing the patient record
CRUD	The identification code of the activity (C: Create, R: Read, U: Update, D: Delete)
Access Control Status	Access control status
SessionID	The identification of the session access
AccessPatient_Warnings	Warning for unusual access
Module Used	The module accessed by the employee

Feature Extraction

To develop the anomaly detection model, including the role classification model, some features were extracted. Each log entry represents a single transaction for a user. To analyze the user activity, the logs from each user were consolidated into a particular period. Every single activity of Doctor A would represent meaningless data points that would be difficult to analyze separately. However, by observing several activities of Doctor A for a particular period, it is easier to perform the anomaly detection task. We processed the log data into 24-hour blocks so that an instance represents the cumulative activity of a user in a single day. As a result, 25,151 instances were extracted from the raw logs, with 24,223 of them being considered normal, 585 considered nonmalicious anomaly, and 343 labeled malicious. Any access that was not for the intention of providing therapeutic functions constitutes malicious access [59]. Therefore, in the logs, malicious data represent all instances that had at least one malicious log access in a single day. The normal data represent all instances in which all of the accesses to the logs are legitimate, and the nonmalicious anomaly data represent the instances that had at least one abnormal log access, but none of them was malicious. These instances were then transformed into features for malicious access detection. Table 7 shows the features extracted from the data set.

Table 7. Features and their related descriptions.

Name of feature	Description
Number of create	Number of created transactions in a single day
Number of reads	Number of read transactions in a single day
Number of updates	Number of updated transactions in a single day
Number of deletes	Number of deleted transactions in a single day
Number of patient records	Number of accesses to patient records in a single day
Number of unique patients	Number of unique patients' records accessed in a single day
Number of modules	Number of the types of modules in the information system accessed in a single day
Number of report modules	Number of transactions in the report modules in a single day
Number of finance modules	Number of finance modules accessed in a single day
Number of patient modules	Number of transactions in the patient module in a single day
Number of lab modules	Number of transactions in the laboratory module in a single day
Number of pharmacy modules	Number of transactions in the pharmacy module in a single day
Number of outside access	Number of transactions from outside the hospital network in a single day
Number of other browsers	Number of browser types used in a single day
Number of Chrome	Number of Chrome uses in a single day
Number of Internet Explorer	Number of Internet Explorer uses in a single day
Number of Safari	Number of Safari uses in a single day
Number of Firefox	Number of Firefox uses in a single day
Number of browsers	Number of other browsers used in a single day

Performance Evaluation for Malicious Detection

For malicious access detection, several measurements, including precision, recall, and F-measures, were identified and used to evaluate the performance. All measurements were calculated based on the confusion matrix displayed in Table 8.

Table 8. Confusion matrix.

Actual	Predicted	
	Malicious	Nonmalicious
Malicious	True positive	False negative
Nonmalicious	False positive	True negative

True positive (TP) and true negative (TN) are the respective number of features that were correctly predicted. TP represents the malicious data that were correctly predicted as malicious, whereas TN represents the nonmalicious data that were correctly predicted as nonmalicious. False positive (FP), also often called the type I error, is the number of nonmalicious data incorrectly predicted as malicious, and false negative (FN), or the type II error, represents the malicious data incorrectly predicted as nonmalicious. The following are the formulas for each measurement:

$$\text{Precision}=TP/TP+FP \quad \textbf{(1)}$$

$$\text{Recall}=TP/TP+FN \quad \textbf{(2)}$$

$$F1=2\times([\text{precision}\times\text{recall}]/[\text{precision}+\text{recall}]) \quad \textbf{(3)}$$

$$F_\beta=(1+\beta^2)(\text{precision}\times\text{recall})/([\beta^2\times\text{precision}]+\text{recall}) \quad \textbf{(4)}$$

Equation 3 is the standard F-score formula where precision and recall have the same weight. If we want to give heavier weight to either precision or recall, we can use equation 4. For any positive real number β, equation 4 is the general F-measure formula where recall is considered to be more important than precision by a weight of β [60]. In this work, we also used the $F_{0.5}$-score and F_2-score. $F_{0.5}$-score means that precision is considered to be two times more important than recall. In contrast, F_2-score means that recall is considered to be two times more important than precision. To compute the $F_{0.5}$-score, the β value was substituted with 0.5, whereas the F_2-score was calculated by replacing the β value with 2.

Usually, automatic malicious behavior detection is used as a filter to narrow down the data for further manual investigation. In this case, high recall is preferred so that most of the actual malicious access will not be missed. Therefore, F_2 is the better measure for this case. However, if we want to use the result

from automatic malicious behavior detection as the final decision without further manual investigation, high precision is preferred over high recall. By using a high-precision method, almost all of the banned accesses are actually malicious. In contrast, if we use an algorithm that prefers high recall as the final decision-maker, we may ban some legitimate accesses that are mistakenly considered fraudulent. In this case, F_1 is the better measure. However, the latter case is rarely applied in the real world since malicious behavior detection is mainly used for a decision support system before further manual investigation.

In this study, we used the logs from January to July as training data, whereas data from August to December were used for testing. The training data were used to train the role classification model, and then this model was used to detect anomalies based on the two proposed approaches. The training data contained a total of 14,558 instances with 13,977 normal instances, 339 nonmalicious anomaly instances, and 242 malicious instances. The testing data consisted of a total of 10,593 instances, with 10,246 normal instances, 246 nonmalicious anomaly instances, and 101 malicious instances.

Experimental Results

The simulation results are summarized in Table 9 and Table 10. Table 9 shows the anomaly detection results from the first stage of two-stage malicious detection. Based on the result, the DT algorithm obtained the best precision (0.655), while the best recall was achieved by SVM (0.977). However, the best F1-score was obtained by RF (0.775). Therefore, the result that was used in the second stage was that obtained from the RF method.

Table 9. Anomaly detection results from the first step of two-stage malicious detection.

Classifier	Precision	Recall	F_1
Multinomial NB[a]	0.256	0.107	0.151
Bernouilli NB	0.256	0.824	0.391
Gaussian NB	0.256	0.618	0.362
KNN[b]	0.634	0.890	0.740
NN[c]	0.651	0.941	0.770
LR[d]	0.242	0.976	0.387
RF[e]	0.662	0.934	0.775
DT[f]	0.665	0.924	0.773
SVM[g]	0.250	0.977	0.399

[a]NB: naive Bayes.

[b]KNN: k-nearest neighbor.

[c]NN: neural network.

[d]LR: logistic regression.

[e]RF: random forest.

[f]DT: decision tree.

[g]SVM: support vector machine.

Table 10 shows the malicious detection results using three approaches. The two-class approach tended to have better performance than the other two approaches. The best precision in the two-stage approach was obtained by LR with a perfect value (1.00), and KNN also had perfect precision in the three-class approach. Three classifiers (RF, DT, and SVM) in the two-class approach achieved the best precision of 0.998.

Furthermore, the best recall was obtained by NN, RF, and DT in the three-classes approach, and by Bernoulli NB and Gaussian NB in both the three-class and two-class approaches. The best F_1 score was obtained by LR in the two-stage approach, SVM in the three-class approach, and Bernoulli NB in the two-class approach. The highest $F_{0.5}$ score was achieved by LR, SVM, and Bernoulli NB in the two-stage, three-class, and two-class approach, respectively. Furthermore, NN and DT achieved the best F_2 score in the two-stage approach, SVM had the best F_2 score in the three-class approach, and Bernoulli NB had the best F_2 score in the two-class approach. Overall, Bernoulli NB with the two-class approach achieved the best F_1, $F_{0.5}$, and F_2 scores.

Table 10. Malicious detection results using three approaches.

Classifier	Two stage	Three classes	Two classes
Multinomial NB[a]			
Precision	0.974	0.931	0.958
Recall	0.752	0.802	0.831
F_1	0.849	0.862	0.890
$F_{0.5}$	0.920	0.902	0.930
F_2	0.788	0.825	0.854
Bernoulli NB			
Precision	0.977	0.824	0.997
Recall	0.832	0.881	0.881
F_1	0.898	0.852	0.935
$F_{0.5}$	0.944	0.835	0.971
F_2	0.857	0.869	0.902
Gaussian NB			
Precision	0.977	0.695	0.994
Recall	0.832	0.881	0.881
F_1	0.898	0.777	0.934
$F_{0.5}$	0.944	0.726	0.969
F_2	0.857	0.836	0.901
KNN[b]			
Precision	0.757	1.000	0.997
Recall	0.832	0.703	0.702
F_1	0.792	0.826	0.824
$F_{0.5}$	0.771	0.922	0.920
F_2	0.816	0.747	0.746
NN[c]			
Precision	0.977	0.977	0.998
Recall	0.842	0.851	0.851
F_1	0.904	0.910	0.919
$F_{0.5}$	0.947	0.949	0.965
F_2	0.866	0.874	0.877
LR[d]			
Precision	1.000	0.966	0.998
Recall	0.832	0.842	0.841
F_1	0.908	0.899	0.913
$F_{0.5}$	0.961	0.938	0.962
F_2	0.861	0.864	0.868
RF[e]			
Precision	0.966	0.966	0.998
Recall	0.842	0.832	0.831

Classifier	Two stage	Three classes	Two classes
F_1	0.899	0.894	0.907
$F_{0.5}$	0.938	0.935	0.959
F_2	0.864	0.855	0.860
DT[f]			
Precision	0.977	0.954	0.998
Recall	0.842	0.822	0.841
F_1	0.904	0.883	0.913
$F_{0.5}$	0.947	0.924	0.962
F_2	0.866	0.845	0.868
SVM[g]			
Precision	0.988	0.978	0.998
Recall	0.832	0.861	0.861
F_1	0.903	0.916	0.924
$F_{0.5}$	0.952	0.952	0.967
F_2	0.859	0.882	0.885

[a]NB: naive Bayes.
[b]KNN: k-nearest neighbor.
[c]NN: neural network.
[d]LR: logistic regression.
[e]RF: random forest.
[f]DT: decision tree.
[g]SVM: support vector machine.

Discussion

Principal Findings

The main purpose of this study was to identify and assess the effectiveness of AI methods and suitable health care staff–generated security practice data for measuring the security practice of health care staff in the context of big data. The main review findings are shown in Table 11. Eighteen studies met the inclusion and exclusion criteria. Recently, a related review for countermeasures against internal threats in health care also identified five machine learning methods that were fit for such measures [30]. This suggests that the adoption of AI methods for modeling and analyzing health care professional–generated security practice data is still an emerging topic of academic interest.

Table 11. Principal findings of the review.

Category	Most used
Algorithms	KNN[a] and Bayesian networks
Features	User IDs, patient IDs, device ID, date and time, location, route, and actions
Data sources	EHR[b] and network logs
Security failures	Anomaly detection
Performance methods	True positive, false positive, false negative, ROC[c] curve, AUC[d]
Data format	CSV[e]
Nature of data sources	Real data logs
Ground truth	Similarity measures and observed data
Privacy preserving approaches	Tokenization and deidentification

[a]KNN: k-nearest neighbor.

[b]EHR: electronic health record.

[c]ROC: receiver operating characteristic.

[d]AUC: area under the receiver operating characteristic curve.

[e]CSV: comma separated value.

AI Methods

As shown in Tables 2 and 11, various algorithms were identified in the study, but the most used methods were KNN and NB algorithms. KNN is a supervised learning–based classification algorithm [44], which learns from labeled data. The KNN then tries to classify unlabeled data items based on the category of the majority of the most similar training data items known as K. The similarity between two data items in KNN can be determined according to the Euclidean distance of the various respective feature vectors of the data items [61]. NB is a probabilistic classifier algorithm based on the assumption that related pairs of features used for determining an outcome are independent of each other and equal [44]. There are two commonly used methods of NB for classifying text: multivariant Bernoulli and multinomial models. KNN and NB algorithms have been more commonly used based on their comparatively higher detection accuracy. For instance, in an experimental assessment of KNN and NB for security countermeasures of internal threats in health care, both models showed over 90% accuracy with NB having a slight advantage over KNN (94% vs 93%). In a related study [30], the KNN method was found to have a higher detection rate with high TP rates and low FP rates.

The major issue with KNN in the context of health care staff security–generated data is the lack of appropriate labeled data [24,53,62]. Within the health care setting, emergencies often dictate needs. In such situations, broader access to resources is normally allowed, making it challenging for reliable labeled data [24,53,62]. Therefore, in adopting KNN for empirical studies, the availability of appropriate labeled data should be considered; however, in the absence of labeled data, unsupervised clustering methods such as K-means clustering could also be considered [26].

Input Data

The input data that were mostly used in the reviewed studies include EHR logs and network data. Yeng et al [4] analyzed observational measures toward profiling health care staff security practices, and also identified various sources, including EHR logs, browser history, network logs, and patterns of keystroke dynamics [4]. Most EHR systems use an emergency access control mechanism known as “break-the-glass” or self-authorization” [1,2]. This enables health care staff to access patients’ medical records during emergency situations without passing through conventional procedures for access authorization. A study [2] into access control methods in Norway revealed that approximately 50% of 100,000 patient records were accessed by 12,298 health care staff (representing approximately 45% of the users) through self-authorization. In such a scenario, EHR remains a vital source for analyzing deviations of required health care security practices.

Ground truth refers to the baseline, which is often used for training the algorithms [63]. The detection efficiency of the algorithms can be negatively impacted if the accuracy of the ground truth is low. As shown in Table 11, various methods—such as similarity measures, observed data, and historical methods—have been used. A similarity measure compares security practices with those of other health care professionals who have similar security practices. The observed measure is a control approach of obtaining the ground truth, whereby some users were observed to conduct their security practices under supervised, required settings [49]. However, the historical data have mainly relied on past records with a trust that the data are sufficiently reliable for the training set. These methods can be assessed for adoption in related studies.

Features and Data Format

EHRs contain most of the features that were identified in this review, as shown in Table 3. Features such as patient ID, actions, and user ID are primary features in EHR logs. The users’ actions

such as deletion, inserting, and updating, and various routes such as diagnosis, prescriptions, and drug dispensing can be tracked in EHR logs [2]. Guided with these findings, the simulated logs contained such attributes and features. Additionally, the simulation of the attributes of logs was also based on the security requirements of the EHRs of Norway [3,4,64,65]. Eventually, a total of 21 attributes and 19 features were included in the simulated logs, as shown in Tables 6 and 7, respectively.

Security Failures and Privacy-Preserving Log Analysis

The application of AI methods to analyze big data generated by health care professional security practice is a reactive approach. With such approaches, the primary aim is to determine deviations or outliers and maliciousness in health care security practices. Anomaly in this work refers to security practices in the access logs that deviate from established security and privacy policies in accessing patient records. For instance, health care workers could be required to access patient records if the health care staff is responsible for the patient throughout their shift and for therapeutic functions. However, it becomes abnormal if the health care staff access patient records outside of their shift. Additionally, if a patient's records are accessed when the patient has not registered for a visit to the hospital, this can also be considered abnormal. Furthermore, if health care staff are accessing patients' records more than usual, this also raises abnormal concerns, although some anomalous access could be for therapeutic purposes and not with ill intentions. However, access that is not for therapeutic functions is described in this work as malicious. A greater proportion of the algorithms were applied for anomaly detection (67%). The detection of anomaly can clearly help in identifying the security practices that deviate from established security policies. However, Rostad and Edsberg [2] found that irregular access to patient records through self-authorization tended to be the normal security practice. An EHR system where a lot of access does not follow the established flow can make it unfeasible to manually track access with malicious intent [2]. Processing that incorporates the detection of malicious access, including intrusion detection, rather than merely detecting outliers could be an effective method of analyzing the security practice in the logs. Therefore, the identified 33% intrusion detections in the review were combined with maliciousness for the simulation since the outcome is to circumvent security requirement in both cases.

Privacy preservation in data mining provides a method to efficiently analyze data while shielding the identifications of the data subjects in a way that respects their right to privacy [66]. In the review, tokenization [43], deidentification [45], and removal of medical information [24] were some methods adopted to preserve privacy. The application of privacy-preserving methods in analyzing log data is crucial since health care data are classified among the most sensitive personal data [67]. Additionally, privacy-preserving methods need to be adopted in compliance with various regulations such as the General Data Protection Regulation [68]. Based on these findings from the review, a roadmap was drawn as a framework for empirical analysis of security practice in the big data context.

Research Implication and Practice

In this work, a comprehensive review was performed in security practice analysis, focusing on the use of AI methods to analyze logs of health care staff. Various AI algorithms, data sources, ground truth, features, application domain data file format, and nature of data sources were identified, analyzed, and modeled. To the best of our knowledge, this is the first time such a study has been systematically performed, along with development of a model and practical assessment of the model with simulated logs for future analysis with actual health care logs. In real log analysis, essential privacy measures such as tokenization and deidentification can be adopted.

Based on the review, a concept was established (Figure 3) on how data-driven and AI methods should be adopted to analyze the logs of EHRs in security practice. The concepts (two-stage, two-class, and three-class) were implemented and their performance was assessed with simulated logs. The attributes of the logs were comprehensive based on the review, which is another major contribution of this study. In the space of supervised learning, our findings pinpoint the suitable algorithms and classification approaches that should be adopted for effective analysis of health care security practices.

Overall, the results of the simulation (Tables 9 and 10) showed that it is easier to differentiate between malicious and nonmalicious access than to distinguish between normal and nonmalicious abnormal access, which is mainly evident from the results of the two-stage approach. The performances of all classifiers in the second stage were far better than those in the first stage. This could also explain why the two-class approach was generally better than the two-stage and three-class approaches. Although the simulated data exhibited good performance with these methods, it is important to recognize that simulated data vary from real data; in particular, real data can be noisier and tend to have an adverse impact on a method's performance [25]. In the application of real data in this framework, effective preprocessing must be carried out toward reducing the noise and its related consequences.

Conclusion

Based on the galloping rate of data breaches in health care, HSPAMI was initiated to observe, model, and analyze health care staff security practices. One of the approaches in HSPAMI is the adoption of AI methods for modeling and analyzing health care staff–generated security practice data [4,16]. This study was then performed to identify, assess, and analyze the appropriate AI methods and data sources. Out of 130 articles that were initially identified in the context of human-generated health care data for security measures in health care, 18 articles were found to meet the inclusion and exclusion criteria. After assessment and analysis, various methods such as KNN, NB, and DT were found to have been mainly applied on EHR logs with varying input features of health care staff security practices. A framework was therefore developed and practically assessed with simulated logs based on the review, toward analyzing real EHR logs.

Based on the results, for anomaly detection, DT algorithms obtained the best precision of 0.655, whereas the best recall was

achieved by SVM at 0.977. However, the best F1-score was obtained by RF at 0.775. In brief, three classifiers (RF, DT, and SVM) in the two-class approach achieved the best precision of 0.998. Moreover, for malicious access detection, LR with the two-stage approach and KNN with the three-class approach obtained perfect precision (1.00), and the best recall was obtained by Bernoulli NB and Gaussian NB in both the three-class and two-class approaches with a value of 0.881. Furthermore, the best F_1 score, $F_{0.5}$ score, and F_2 score for malicious access detection were achieved by Bernoulli NB using the two-class approach with values of 0.935, 0.971, and 0.902, respectively. These methods can therefore be used in analyzing health care security practice toward finding incentive measures for information security compliance in the health care sector. This study covered only supervised learning where labeled data were used. Future work is therefore required using unsupervised learning methods in analyzing logs that do not have labeled data.

Conflicts of Interest

None declared.

References

1. Ardagna C, De Capitani di Vimercati S, Foresti S, Grandison T, Jajodia S, Samarati P. Access control for smarter healthcare using policy spaces. Comput Secur 2010 Nov;29(8):848-858. [doi: 10.1016/j.cose.2010.07.001]
2. Rostad L, Edsberg O. A study of access control requirements for healthcare systems based on audit trails from access logs. : IEEE; 2006 Nov 15 Presented at: Annual Computer Security Applications Conference (ACSAC'06); December 2006; Miami Beach, FL p. 11-15. [doi: 10.1109/ACSAC.2006.8]
3. Code of conduct version 6. Direktorate for e-health, Norway. 2020 Dec 15. URL: https://www.ehelse.no/normen/documents-in-english [accessed 2020-12-15]
4. Yeng P, Yang B, Snekkenes E. Observational measures for effective profiling of healthcare staffs' security practices. : IEEE; 2019 Jul 15 Presented at: 43rd Annual Computer Software and Applications Conference (COMPSAC); July 2019; Milwaukee, WI p. 15-19 URL: https://ieeexplore.ieee.org/document/8754403 [doi: 10.1109/compsac.2019.10239]
5. Nweke LO, Yeng P, Wolthusen SD, Yang B. Understanding attribute-based access control for modelling and analysing healthcare professionals' security practices. J Adv Comput Sci Appl 2020;11(2):683-690. [doi: 10.14569/IJACSA.2020.0110286]
6. Baro E, Degoul S, Beuscart R, Chazard E. Toward a literature-driven definition of big data in healthcare. Biomed Res Int 2015;2015:639021. [doi: 10.1155/2015/639021] [Medline: 26137488]
7. Yeng P, Yang B, Snekkenes E. Framework for healthcare security practice analysis, modeling and incentivization. : IEEE; 2019 Dec 09 Presented at: International Workshop on Big Data Analytics for Cyber Threat Hunting; December 2019; Los Angeles, CA p. 9-12 URL: https://ieeexplore.ieee.org/document/9006529 [doi: 10.1109/bigdata47090.2019.9006529]
8. Ramesh A, Kambhampati C, Monson J, Drew P. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004 Sep 01;86(5):334-338 [FREE Full text] [doi: 10.1308/147870804290] [Medline: 15333167]
9. Widup S. 2019 Verizon Data Breach Investigations Report. NIST. 2019 Dec. URL: https://www.nist.gov/system/files/documents/2019/10/16/1-2-dbir-widup.pdf [accessed 2020-11-12]
10. Taylor T. Hackers, Breaches, and the Value of Healthcare Data. SecureLink. 2021 Jun 30. URL: https://www.securelink.com/blog/healthcare-data-new-prize-hackers/ [accessed 2021-12-15]
11. Humer C, Finkle J. Your medical record is worth more to hackers than your credit card. Reuters. 2014 Sep 24. URL: https://www.reuters.com/article/us-cybersecurity-hospitals-idUSKCN0HJ21I20140924 [accessed 2021-08-03]
12. Garrity M. Patient medical records sell for $1K on dark web. Becker's Hospital. 2020 Dec. URL: https://www.beckershospitalreview.com/cybersecurity/patient-medical-records-sell-for-1k-on-dark-web.html [accessed 2021-01-01]
13. Cannoy SD, Salam AF. A framework for health care information assurance policy and compliance. Commun ACM 2010 Mar;53(3):126-131. [doi: 10.1145/1666420.1666453]
14. Safa NS, Sookhak M, Von Solms R, Furnell S, Ghani NA, Herawan T. Information security conscious care behaviour formation in organizations. Comput Secur 2015 Sep;53:65-78. [doi: 10.1016/j.cose.2015.05.012]
15. Whitman M, Fendler P, Caylor J, Baker D. Rebuilding the human firewall. 2005 Sep 05 Presented at: 2nd annual conference on Information security curriculum development; 2005; Kennesaw, Georgia p. 104-106. [doi: 10.1145/1107622.1107646]
16. Yeng PK, Szekeres A, Yang B, Snekkenes EA. Mapping the psychosocialcultural aspects of healthcare professionals' information security practices: systematic mapping study. JMIR Hum Factors 2021 Jun 09;8(2):e17604 [FREE Full text] [doi: 10.2196/17604] [Medline: 34106077]
17. Riggs C. Chapter 8: Firewalls. In: Network perimeter security: building defense in-depth. New York: Auerbach Publications; Oct 27, 2003.
18. Predd J, Pfleeger S, Hunker J, Bulford C. Insiders behaving badly. IEEE Secur Privacy Mag 2008 Jul;6(4):66-70 [FREE Full text] [doi: 10.1109/msp.2008.87]
19. McLeod A, Dolezel D. Cyber-analytics: modeling factors associated with healthcare data breaches. Dec Support Syst 2018 Apr;108:57-68. [doi: 10.1016/j.dss.2018.02.007]

20. McLeod A, Dolezel D. Understanding healthcare data breaches: crafting security profiles. 2018 Aug 16 Presented at: 24th Americas Conference on Information Systems; August 16, 2018; New Orleans, LA p. 16-18 URL: https://dblp.org/rec/conf/amcis/McLeodD18.bib
21. Kwon J, Johnson M. The market effect of healthcare security: do patients care about data breaches? 2015 Jun 22 Presented at: Workshop on the Economics of Information Security; June 22-23, 2015; Netherlands URL: https://scholars.cityu.edu.hk/en/publications/publication(76aa2cc3-dd5d-4f82-9856-a5bf3c2fcc1f).html
22. Yeng PK, Yang B, Snekkenes EA. Healthcare staffs' information security practices towards mitigating data breaches: a literature survey. Stud Health Technol Inform 2019;261:239-245. [Medline: 31156123]
23. Shaban-Nejad A, Michalowski M, Buckeridge DL. Health intelligence: how artificial intelligence transforms population and personalized health. NPJ Digit Med 2018;1:53. [doi: 10.1038/s41746-018-0058-9] [Medline: 31304332]
24. Ziemniak T. Use of machine learning classification techniques to detect atypical behavior in medical applications. : IEEE; 2011 Jun 27 Presented at: Sixth International Conference on IT Security Incident Management and IT Forensics; May 10-12, 2011; Stuttgart, Germany p. 10-12. [doi: 10.1109/IMF.2011.20]
25. Kononenko I, Kukar M. Machine learning and data mining: introduction to principles and algorithms. Sawston, Cambridge, UK: Horwood Publishing Ltd; Sep 10, 2007.
26. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
27. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health 2018;3(4):e000798 [FREE Full text] [doi: 10.1136/bmjgh-2018-000798] [Medline: 30233828]
28. Vihinen M, Samarghitean C. Medical expert systems. Curr Bioinform 2008 Jan 01;3(1):56-65. [doi: 10.2174/157489308783329869]
29. Chandra S, Ray S, Goswami R. Big data security in healthcare: survey on frameworks and algorithms. 2017 Jan 05 Presented at: 2017 IEEE 7th International Advance Computing Conference (IACC); January 5-7, 2017; Hyderabad p. 5-7. [doi: 10.1109/iacc.2017.0033]
30. Walker-Roberts S, Hammoudeh M, Dehghantanha A. A systematic review of the availability and efficacy of countermeasures to internal threats in healthcare critical infrastructure. IEEE Access 2018;6:25167-25177. [doi: 10.1109/access.2018.2817560]
31. Bose B, Avasarala B, Tirthapura S, Chung Y, Steiner D. Detecting insider threats using RADISH: a system for real-time anomaly detection in heterogeneous data streams. IEEE Syst J 2017 Jun;11(2):471-482. [doi: 10.1109/jsyst.2016.2558507]
32. Gafny M, Shabtai A, Rokach L, Elovici Y. Detecting data misuse by applying context-based data linkage. 2010 Sep 21 Presented at: ACM workshop on Insider Threats; October 2010; Chicago, IL. [doi: https://doi.org/10.1145/1866886.1866890]
33. Chen Y, Nyemba S, Zhang W, Malin B. Specializing network analysis to detect anomalous insider actions. Secur Inform 2012 Feb 27;1:5 [FREE Full text] [doi: 10.1186/2190-8532-1-5] [Medline: 23399988]
34. Islam MS, Hasan MM, Wang X, Germack HD, Noor-E-Alam M. A systematic review on healthcare analytics: application and theoretical perspective of data mining. Healthcare (Basel) 2018 May 23;6(2):54 [FREE Full text] [doi: 10.3390/healthcare6020054] [Medline: 29882866]
35. Gheyas IA, Abdallah AE. Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis. Big Data Anal 2016 Aug 30;1(1):1-14. [doi: 10.1186/s41044-016-0006-0]
36. Khraisat A, Gondal I, Vamplew P, Kamruzzaman J. Survey of intrusion detection systems: techniques, datasets and challenges. Cybersecur 2019 Jul 17;2(1):20. [doi: 10.1186/s42400-019-0038-7]
37. Yeng PK, Nweke LO, Woldaregay AZ, Yang B, Snekkenes EA. Data-driven and artificial intelligence (AI) approach for modelling and analyzing healthcare security practice: a systematic review. Cham: Springer; 2021 Presented at: Intelligent Systems and Applications. IntelliSys 2020; August 25, 2020; Online. [doi: 10.1007/978-3-030-55180-3_1]
38. Kitchenham B, Pretorius R, Budgen D, Pearl Brereton O, Turner M, Niazi M, et al. Systematic literature reviews in software engineering – a tertiary study. Inf Softw Technol 2010 Aug;52(8):792-805. [doi: 10.1016/j.infsof.2010.03.006]
39. Booth A, Sutton A, Papaioannou D. Systematic approaches to a successful literature review. Thousand Oaks, CA: Sage Publications; 2016.
40. Khan R, Khan S. A preliminary structure of software security assurance model. 2018 Presented at: ICGSE '18: Proceedings of the 13th International Conference on Global Software Engineering; May 2018; Gothenburg, Sweden. [doi: 10.1145/3196369.3196385]
41. Petersen K, Vakkalanka S, Kuzniarz L. Guidelines for conducting systematic mapping studies in software engineering: An update. Inf Soft Technol 2015 Aug;64:1-18. [doi: 10.1016/j.infsof.2015.03.007]
42. PRISMA statement. PRISMA. 2018. URL: http://www.prisma-statement.org/ [accessed 2020-08-15]
43. Boddy AJ, Hurst W, Mackay M, Rhalibi AE. Density-based outlier detection for safeguarding electronic patient record systems. IEEE Access 2019;7:40285-40294. [doi: 10.1109/access.2019.2906503]
44. García Adeva JJ, Pikatza Atxa JM. Intrusion detection in web applications using text mining. Engineer Appl Artif Intel 2007 Jun;20(4):555-566. [doi: 10.1016/j.engappai.2006.09.001]

45. Gupta S, Hanson C, Gunter C, Frank M, Liebovitz D, Malin B. Modeling and detecting anomalous topic access. 2013 Jun 16 Presented at: IEEE International Conference on Intelligence and Security Informatics; June 4-7, 2013; Seattle, WA p. 4-7 URL: https://ieeexplore.ieee.org/document/6578795 [doi: 10.1109/isi.2013.6578795]
46. Chen Y, Nyemba S, Malin B. Detecting anomalous insiders in collaborative information systems. IEEE Trans Dependable and Secure Comput 2012 May;9(3):332-344. [doi: 10.1109/tdsc.2012.11]
47. Chen Y, Malin B. Detection of anomalous insiders in collaborative environments via relational analysis of access logs. CODASPY 2011;2011:63-74 [FREE Full text] [doi: 10.1145/1943513.1943524] [Medline: 25485309]
48. Amálio N, Spanoudakis G. From monitoring templates to security monitoring and threat detection. 2018 Presented at: Second International Conference on Emerging Security Information, Systems and Technologies; 2008; Esterel, France p. 25-31. [doi: 10.1109/securware.2008.58]
49. Wesołowski TE, Porwik P, Doroz R. Electronic health record security based on ensemble classification of keystroke dynamics. Appl Artif Intel 2016 Jul 21;30(6):521-540. [doi: 10.1080/08839514.2016.1193715]
50. Pierrot D, Harbi N, Darmont J. Hybrid intrusion detection in information systems. 2016 Presented at: International Conference on Information Science and Security (ICISS); December 19-22, 2016; Pattaya p. 19-22. [doi: 10.1109/icissec.2016.7885857]
51. Menon AK, Jiang X, Kim J, Vaidya J, Ohno-Machado L. Detecting inappropriate access to electronic health records using collaborative filtering. Mach Learn 2014 Apr 01;95(1):87-101 [FREE Full text] [doi: 10.1007/s10994-013-5376-1] [Medline: 24683293]
52. Tchakoucht T, Ezziyyani M, Jbilou M, Salaun M. Behavioral approach for intrusion detection. : IEEE; 2016 Jul 17 Presented at: IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA); November 17-20, 2015; Marrakech, Morocco p. 17-20.
53. Boddy A, Hurst W, Mackay M, Rhalibi A. A hybrid density-based outlier detection model for privacy in electronic patient record system. 2019 Presented at: International Conference on Information Management (ICIM); March 2019; Cambridge, UK p. 24-27. [doi: 10.1109/infoman.2019.8714701]
54. Li X, Xue Y, Malin B. Detecting anomalous user behaviors in workflow-driven web applications. : IEEE; 2013 Feb 10 Presented at: IEEE 31st Symposium on Reliable Distributed Systems; October 2013; Irvine, CA p. 8-11 URL: https://ieeexplore.ieee.org/document/6424834 [doi: 10.1109/srds.2012.19]
55. Costante E, Fauri D, Etalle S, Hartog J, Zannone N. A hybrid framework for data loss prevention and detection. : IEEE; 2016 Presented at: 2016 IEEE Security and Privacy Workshops (SPW); August 4, 2016; San Jose, CA. [doi: 10.1109/spw.2016.24]
56. Zhang H, Mehotra S, Liebovitz D, Gunter CA, Malin B. Mining deviations from patient care pathways via electronic medical record system audits. ACM Trans Manage Inf Syst 2013 Dec;4(4):1-20. [doi: 10.1145/2544102]
57. SIRAJ A, VAUGHN RB, BRIDGES SM. Decision making for network health assessment in an intelligent intrusion detection system architecture. Int J Info Tech Dec Mak 2011 Nov 20;03(02):281-306. [doi: 10.1142/s0219622004001057]
58. Asfaw B, Bekele D, Eshete B, Villafiorita A, Weldemariam K. Host-based anomaly detection for pervasive medical systems. 2010 Presented at: Fifth International Conference on Risks and Security of Internet and Systems (CRiSIS); October 2010; Montreal, QC, Canada. [doi: 10.1109/crisis.2010.5764923]
59. e-helse D. Implementation of GDPR in health care sector in Norway. Directorate of e-health. 2019. URL: https://www.ehelse.no/personvern-og-informasjonssikkerhet/implementation-of-gdpr-in-health-care-sector-in-norway#:~:text=GDPR%20was%20adopted%20in%20the,Norway%20on%2020th%20July%202018 [accessed 2021-01-15]
60. Fauzi MA, Bours P. Ensemble method for sexual predators identification in online chats. 2020 Jun 04 Presented at: 2020 8th International Workshop on Biometrics and Forensics (IWBF); April 29-30, 2020; Porto, Portugal. [doi: 10.1109/iwbf49977.2020.9107945]
61. Yeng P, Woldaregay A, Hartvigsen G. K-CUSUM: cluster detection mechanism in EDMON. 2019 Presented at: SHI: 7th Scandinavian Conference on Health Informatics; November 2019; Oslo, Norway p. 12-13 URL: https://ep.liu.se/konferensartikel.aspx?series=ecp&issue=161&Article_No=24
62. Gates C, Li N, Xu Z, Chari S, Molloy I, Park Y. Detecting insider information theft using features from file access logs. In: Kutyłowski M, Vaidya J, editors. Computer Security - ESORICS 2014. Lecture Notes in Computer Science, vol 8713. Cham: Springer; 2014:383-400.
63. Smyth P, Fayyad U, Burl M, Perona P, Baldi P. Inferring ground truth from subjective labelling of venus images. Cambridge, MA: MIT Press; 1996 Presented at: NIPS'94: 7th International Conference on Neural Information Processing Systems; January 1994; Denver, CO p. 1085-1092 URL: https://resolver.caltech.edu/CaltechAUTHORS:20150305-153627706
64. Yeng PK, Fauzi MA, Yang B. Workflow-based anomaly detection using machine learning on electronic health records' logs: a comparative study. 2021 Jun 23 Presented at: 2020 International Conference on Computational Science and Computational Intelligence (CSCI); December 16-18, 2020; Las Vegas, NV p. 753-760. [doi: 10.1109/csci51800.2020.00143]
65. Yeng PK, Fauzi MA, Yang B. Comparative analysis of machine learning methods for analyzing security practice in electronic health records' logs. 2020 Presented at: 2020 IEEE International Conference on Big Data (Big Data); December 10-13, 2020; Virtual p. 3856-3866. [doi: 10.1109/BigData50022.2020.9378353]
66. Agrawal R, Srikant R. Privacy-preserving data mining. 2000 Presented at: ACM SIGMOD International Conference on Management of Data; May 15-18, 2000; Dallas, TX. [doi: 10.1145/342009.335438]

67. ISO 27799: 2016 Health informatics — Information security management in health using ISO/IEC 27002. ISO. 2016. URL: https://www.iso.org/standard/62777.html [accessed 2020-12-20]
68. Pseudonymization. Imperva. 2019. URL: https://www.imperva.com/data-security/compliance-101/pseudonymization/ [accessed 2020-12-20]

Abbreviations

AI: artificial intelligence
CIA: confidentiality, integrity, and availability
DT: decision tree
EHR: electronic health record
FN: false negative
FP: false positive
HSPAMI: Healthcare Security Practice Analysis, Modeling, and Incentivization
LR: logistic regression
NB: naïve Bayes
NN: neural network
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF: random forest
SVM: support vector machine
TN: true negative
TP: true positive

Edited by R Kukafka, G Eysenbach; submitted 09.04.20; peer-reviewed by B Brumen, A Moayedikia, A ., J Ropero, R Sutton, F Alvarez-Lopez; comments to author 08.06.20; revised version received 02.02.21; accepted 28.09.21; published 22.12.21.

Please cite as:
Yeng PK, Nweke LO, Yang B, Ali Fauzi M, Snekkenes EA
Artificial Intelligence–Based Framework for Analyzing Health Care Staff Security Practice: Mapping Review and Simulation Study
JMIR Med Inform 2021;9(12):e19250
URL: https://medinform.jmir.org/2021/12/e19250
doi:10.2196/19250
PMID:34941549

Viewpoint

Can Real-time Computer-Aided Detection Systems Diminish the Risk of Postcolonoscopy Colorectal Cancer?

Mariusz Madalinski[1*], DPhil, MD; Roger Prudham[2*], MBBS, FRCP

[1]Northern Care Alliance, Royal Oldham Hospital, Oldham, United Kingdom

[2]Northern Care Alliance, Bury, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Mariusz Madalinski, DPhil, MD
Northern Care Alliance
Royal Oldham Hospital
Rochdale Rd
Oldham, OL1 2JH
United Kingdom
Phone: 44 01616240420
Email: mariusz.madalinski@googlemail.com

Abstract

The adenoma detection rate is the constant subject of research and the main marker of quality in bowel cancer screening. However, by improving the quality of endoscopy via artificial intelligence methods, all polyps, including those with the potential for malignancy, can be removed, thereby reducing interval colorectal cancer rates. As such, the removal of all polyps may become the best marker of endoscopy quality. Thus, we present a viewpoint on integrating the computer-aided detection (CADe) of polyps with high-accuracy, real-time colonoscopy to challenge quality improvements in the performance of colonoscopy. Colonoscopy for bowel cancer screening involving the integration of a deep learning methodology (ie, integrating artificial intelligence with CADe systems) has been assessed in an effort to increase the adenoma detection rate. In this viewpoint, a few studies are described, and their results show that CADe systems are able to increase screening sensitivity. The detection of adenomatous polyps, which are associated with a potential risk of progression to colorectal cancer, and their removal are expected to reduce cancer incidence and mortality rates. However, so far, artificial intelligence methods do not increase the detection of cancer or large adenomatous polyps but contribute to the detection of small precancerous polyps.

(JMIR Med Inform 2021;9(12):e25328) doi:10.2196/25328

KEYWORDS

artificial intelligence; colonoscopy; adenoma; real-time computer-aided detection; colonic polyp

Introduction

Adenomatous polyps are associated with a potential risk of progression to colorectal cancer (CRC). The adenoma detection rate (ADR) is regarded as an important marker of the quality of inspection in colonoscopy. The identification and removal of adenomatous polyps are considered to be important in CRC prevention [1,2]. More recently, computer-aided detection (CADe) tools that incorporate a 3D fully convolutional network have been developed to aid with colonoscopy screening for CRC. Deep learning methodologies, whereby a programmer teaches a computer which features to focus on, have been developed, thus allowing artificial intelligence (AI) to be integrated during colonoscopy [3,4].

CADe Tools for Colonic Cancer: The Studies

Repici et al [3] have presented results on their evaluation of the efficacy of integrating the CADe of colonic polyps with high-accuracy, real-time colonoscopy. This provides a unique opportunity to obtain real-time feedback for informing an endoscopist about the quality of a live endoscopy.

In Repici et al's [3] study, 685 individuals were randomized, and the authors reported a significantly higher ADR in the CADe group. This appears to confirm the findings of Wang et al [4], who enrolled 1058 patients into their first prospective randomized controlled trial. Both studies reported a significantly higher mean number of adenomas and nonpolypoid lesion

detection rate in the CADe group than those in the control group [3,4].

A real-time automatic detection system that uses deep neural networks was trialed in Italy, and it achieved a high ADR (CADe group: 54.8%; control group: 40.4%) [3]. However, a much lower ADR was reported for both groups (CADe group: 29.1%; control group: 20.3%) in Wang et al's [4] study, but the mean age of the participants in this Chinese study was 49.94 years (SD 13.79 years) in the control group and 51.07 years (SD 13.15 years) in the CADe group [4]. This may also be explained by the observation that the overall prevalence of adenomas and CRC is lower in mainland China than in Europe and the United States [5]. Comparing these studies is difficult however, as in the Repici et al [3] study, the patients' mean age was considerably higher (mean 61.32 years, SD 10.2 years). In their study, a significantly higher number of diminutive adenomas and adenomas that were 6 to 9 mm in diameter were detected in the CADe group, regardless of the adenomas' location or morphology [3]. In the Wang et al [4] study, CADe helped to significantly increase the detection of adenomas in colonic segments (ie, from the hepatic flexure to the rectosigmoid junction), but the CADe technology appeared to be the most effective at detecting adenomas in the transverse colon. A further analysis revealed that the higher ADR in the CADe group was mainly due to an increase in the detection of diminutive adenomas; there were no significant differences among large ADRs [4].

Recently, a Chinese cross-sectional study [5] reported a higher ADR for the proximal colon compared to that for the distal colon, but this difference was not observed in the Wang et al study [4]. However, this difference was observed by Repici et al's [3] team. The ratio of precancerous polyps located in the proximal colon to precancerous polyps in the distal colon is another suggested measure of performance that may be used to confirm the high quality of a clearing colonoscopy [6].

Repici et al's [3] Study Limitations

The six experienced endoscopists in the Repici et al [3] study had over 2000 screening colonoscopies under their belts. We do not know if more experienced endoscopists—those who have performed more than 10,000 colonoscopies—would confirm Repici et al's [3] results. Moreover, the endoscopists were required to adhere to a minimum of 6 minutes for inspection; their mean withdrawal time was around 7 minutes [3] (the withdrawal time was a little shorter in Wang et al's [4] study).

The endoscopists' withdrawal techniques did not meet the criteria for aspirational withdrawal time (≥10 minutes) that are present in the European Society for Gastroenterology guidelines [1] and the British Society of Gastroenterology guidelines [2]. There is evidence that a shorter withdrawal time is associated with a lower ADR and a higher incidence of postcolonoscopy CRC and that a longer withdrawal time increases the ADR [1,2]. The exact mechanism by which withdrawal time impacts the risk of postcolonoscopy CRC and its impacts on the ADR are not well known, but we can hypothesize that withdrawal time affects careful colonic mucosal inspection.

The Future of Bowel Cancer Screening

Endoscopists' withdrawal techniques and specified right colon withdrawal times correlate with higher levels of polyp detection [7]. Therefore, a considerable challenge lies ahead of those who wish to use the detection all polyps (via AI methods) as a new independent marker. Further research is needed to determine whether this marker is more optimal than the advised aspirational withdrawal time (≥10 minutes) in current colonoscopy guidelines or the ADR. Additionally, other interesting questions that have arisen are whether the withdrawal time is a better marker than the ADR and whether these markers are surrogate markers for the detection of all polyps that are monitored via AI. Originally, the ADR was defined as the percentage of patients aged ≥50 years who underwent primary screening colonoscopy for the first time and had 1 or more conventional adenomas [1,2].

The adenoma miss rate varies among endoscopists who achieve the same ADRs, and a significant difference in adenoma miss rates has been reported even among endoscopists who achieve high ADRs [8]. A reduction in the number of all colonic adenomas may be recognized as a complementary benchmark of cancer protection after clearing colonoscopies. Therefore, we assume that the removal of all polyps with the potential for carcinogenesis comprises an independent marker of quality that is relevant to clearing colonoscopies, and AI may be helpful for assessing this goal. Thus, as a support for endoscopists who have not developed the highest quality skills, AI creates a new opportunity, especially after the end of colonoscopy training.

Further studies are required to determine whether AI is of benefit to endoscopists who are more experienced than those in Repici et al's [3] study. Our personal experience reveals that using AI results in the increased incidence of the overdiagnosis of polyps with little or no malignant potential. It is important to not accept as a given that the utility offered by AI-assisted colonoscopy in detecting diminutive polyps is of definite value overall. It is possible that as AI-assisted colonoscopy increases the number of diminutive polyps that are detected, the time taken to complete a colonoscopy also increases, as these polyps must be inspected and removed. This may in turn increase the costs associated with colonoscopy. Within health economies that are constrained by limited resources, AI-assisted colonoscopy may have the unintended consequence of reducing the amount of benefits that are provided to the population as a whole by reducing access to colonoscopy. Long-term outcome studies must be conducted to determine how beneficial this new technology may be, regardless of how exciting it appears to be at first glance.

Conclusion

So far, we know that AI methods do not increase the detection of large adenomas or cancer. The contribution of small adenomas, which have been increasingly detected via AI-assisted colonoscopy, to future CRC risk is debatable.

Conflicts of Interest

None declared.

References

1. Rees CJ, Gibson ST, Rutter MD, Baragwanath P, Pullan R, Feeney M, British Society of Gastroenterology, the Joint Advisory Group on GI Endoscopy, the Association of Coloproctology of Great Britain and Ireland. UK key performance indicators and quality assurance standards for colonoscopy. Gut 2016 Dec;65(12):1923-1929 [FREE Full text] [doi: 10.1136/gutjnl-2016-312044] [Medline: 27531829]
2. Kaminski MF, Thomas-Gibson S, Bugajski M, Bretthauer M, Rees CJ, Dekker E, et al. Performance measures for lower gastrointestinal endoscopy: a European Society of Gastrointestinal Endoscopy (ESGE) Quality Improvement Initiative. Endoscopy 2017 Apr;49(4):378-397 [FREE Full text] [doi: 10.1055/s-0043-103411] [Medline: 28268235]
3. Repici A, Badalamenti M, Maselli R, Correale L, Radaelli F, Rondonotti E, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020 Aug;159(2):512-520.e7. [doi: 10.1053/j.gastro.2020.04.062] [Medline: 32371116]
4. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut 2019 Oct;68(10):1813-1819 [FREE Full text] [doi: 10.1136/gutjnl-2018-317500] [Medline: 30814121]
5. Wong MC, Ding H, Wang J, Chan PS, Huang J. Prevalence and risk factors of colorectal cancer in Asia. Intest Res 2019 Jul;17(3):317-329 [FREE Full text] [doi: 10.5217/ir.2019.00021] [Medline: 31085968]
6. Madalinski M. Ratio of right-sided to left-sided dysplastic colonic polyps is a valid key performance indicator. United European Gastroenterol J 2018 Oct;6(Supplement 1):A192.
7. Rex DK, Schoenfeld PS, Cohen J, Pike IM, Adler DG, Fennerty MB, et al. Quality indicators for colonoscopy. Gastrointest Endosc 2015 Jan;81(1):31-53. [doi: 10.1016/j.gie.2014.07.058] [Medline: 25480100]
8. Aniwan S, Orkoonsawat P, Viriyautsahakul V, Angsuwatcharakon P, Pittayanon R, Wisedopas N, et al. The secondary quality indicator to improve prediction of adenoma miss rate apart from adenoma detection rate. Am J Gastroenterol 2016 May;111(5):723-729. [doi: 10.1038/ajg.2015.440] [Medline: 26809333]

Abbreviations

ADR: adenoma detection rate
AI: artificial intelligence
CADe: computer-aided detection
CRC: colorectal cancer

Edited by C Lovis; submitted 27.10.20; peer-reviewed by J Rayapudi, S Pang; comments to author 16.01.21; revised version received 26.07.21; accepted 23.09.21; published 24.12.21.

Please cite as:
Madalinski M, Prudham R
Can Real-time Computer-Aided Detection Systems Diminish the Risk of Postcolonoscopy Colorectal Cancer?
JMIR Med Inform 2021;9(12):e25328
URL: https://medinform.jmir.org/2021/12/e25328
doi:10.2196/25328
PMID:34571490

Original Paper

Text Mining of Adverse Events in Clinical Trials: Deep Learning Approach

Daphne Chopard[1], MSc; Matthias S Treder[1], PhD; Padraig Corcoran[1], PhD; Nagheen Ahmed[2], BSc; Claire Johnson[2], BSc; Monica Busse[2], PhD; Irena Spasic[1], PhD

[1]School of Computer Science & Informatics, Cardiff University, Cardiff, United Kingdom

[2]Centre for Trials Research, Cardiff University, Cardiff, United Kingdom

Corresponding Author:
Irena Spasic, PhD
School of Computer Science & Informatics
Cardiff University
Abacws
Senghennydd Road
Cardiff, CF24 4AG
United Kingdom
Phone: 44 2920870032
Email: spasici@cardiff.ac.uk

Abstract

Background: Pharmacovigilance and safety reporting, which involve processes for monitoring the use of medicines in clinical trials, play a critical role in the identification of previously unrecognized adverse events or changes in the patterns of adverse events.

Objective: This study aims to demonstrate the feasibility of automating the coding of adverse events described in the narrative section of the serious adverse event report forms to enable statistical analysis of the aforementioned patterns.

Methods: We used the Unified Medical Language System (UMLS) as the coding scheme, which integrates 217 source vocabularies, thus enabling coding against other relevant terminologies such as the International Classification of Diseases–10th Revision, Medical Dictionary for Regulatory Activities, and Systematized Nomenclature of Medicine). We used MetaMap, a highly configurable dictionary lookup software, to identify the mentions of the UMLS concepts. We trained a binary classifier using Bidirectional Encoder Representations from Transformers (BERT), a transformer-based language model that captures contextual relationships, to differentiate between mentions of the UMLS concepts that represented adverse events and those that did not.

Results: The model achieved a high F1 score of 0.8080, despite the class imbalance. This is 10.15 percent points lower than human-like performance but also 17.45 percent points higher than that of the baseline approach.

Conclusions: These results confirmed that automated coding of adverse events described in the narrative section of serious adverse event reports is feasible. Once coded, adverse events can be statistically analyzed so that any correlations with the trialed medicines can be estimated in a timely fashion.

(JMIR Med Inform 2021;9(12):e28632) doi:10.2196/28632

KEYWORDS

natural language processing; deep learning; machine learning; classification

Introduction

Background

Modern health care is associated with increased costs and broad-reaching variations in care and outcomes across the global population. The provision of evidence-based health care is a critical priority for users, providers, and policy makers alike. The systematic and high-quality conduct of clinical trials is critical for the development of clinical guidance to inform evidence-based practice. Pharmacovigilance and safety reporting are among the most important aspects of the conduct of clinical trials. This is relevant to all clinical trials in which the benefit

or harm must be fully established before any intervention or medicinal product is adopted.

Pharmacovigilance and safety reporting provide the basis for ensuring clinical trial participant safety and good research practice. It involves processes for monitoring the use of medicines or interventions in clinical trials. It has a critical role in the identification of previously unrecognized adverse events or changes in the patterns of adverse events. It is also relevant to the assessment of the risks and benefits of medicines or interventions to determine what action, if any, is needed to improve their safe use.

An adverse event is any untoward medical occurrence in a participant to whom a medicinal product has been administered, including occurrences that are not necessarily caused by or related to the administered product. A serious adverse event (SAE) is any untoward medical occurrence that, at any dose, results in death, is life-threatening, requires inpatient hospitalization or causes prolongation of existing hospitalization, results in persistent or significant disability or incapacity, or comprises a congenital anomaly or birth defect. Early detection of unknown adverse events, reactions, interactions, and an increase in the frequency of (known) adverse events is a key element of the pharmacovigilance and safety process. Provision of up-to-date information on adverse events to health care professionals, researchers, and regulatory bodies contributes to the assessment of benefit, harm, effectiveness, and risk of the intervention, thus advancing their safe, rational, and more effective (including cost-effective) use.

In multicenter noncommercial clinical trials conducted in the United Kingdom, the SAE reporting requirements are detailed in the trial protocol, and the principal investigators at National Health Service sites are responsible for reporting SAEs to the coordinating clinical trial unit (CTU) for an assessment of the seriousness, causality, and expectedness as delegated by the clinical trial sponsor. An SAE report includes an event term and additional signs and symptoms in a narrative. The narrative is reported by a physician during their medical assessment of the event. The report is then reviewed by a central CTU reviewer to assess any potential causal relationship with the trial drug. Each narrative is reviewed as a single report. The narratives are typically received from sites as paper records. These are logged electronically in the safety databases by the CTU pharmacovigilance team for the relevant national competent authorities (eg, the UK Medicines and Health Care Products Regulatory Agency or European Medicines Agency). The reports are searchable on request and subject to appropriate regulatory permissions. There is now a clear recognition of the potential for artificial intelligence in safety case management to identify relationships and signals [1]. Although these approaches may be implemented in commercial settings and within competent authorities, such methods for classifying and categorizing data are not yet standardized or explicit across noncommercial pharmacovigilance settings.

It is possible that the narrative contains additional adverse events or toxicities that are not coded as additional events and are captured in the narrative only. However, there is no mechanism for the detection of safety signals across individual reports or individual trials and, thus, there is no possibility for early detection of worrying trends. This is particularly the case for toxicities for which reconciliation with the clinical database would be advantageous. Such a tool would facilitate the cross-checking of toxicities recorded in the narrative of the SAE form with those recorded in the trial database, which is currently only feasible if automated. Although these approaches may be used in commercial trial settings, they would not always be used in the public domain simply because of the nature of the drug licensing pathway.

This study seeks to use text mining to automatically identify and code adverse events from the narrative sections of SAE reports in clinical trials of investigational medicinal products coordinated by a noncommercial CTU, with the aim of unlocking narrative evidence for further statistical analysis. Although such an analysis is beyond the scope of this study, it would serve to monitor the patterns of adverse events at the cohort level rather than singular adverse events. Owing to their narrative nature, such an analysis cannot be conducted directly on the content of SAE reports.

Related Work

Text mining has been used to identify adverse events from a variety of data sources, including spontaneous reporting systems, medical literature, electronic health records, and user-generated content on the internet [2]. The problem of mining adverse events in text has been approached from different angles. Most commonly, it has been defined as a text classification problem, where a piece of text, either an entire document or its part (eg, an individual sentence), is mapped to ≥1 predefined class that correspond to a type of adverse event or its property. Some approaches target a specific adverse event such as anaphylaxis and perform simple binary classification with respect to the presence of the event considered [3]. Other examples target a range of drugs and use documents that mention them to train a binary classifier with respect to their safety, using an existing watch list of drugs that have an active safety alert posted on the US Food and Drug Administration website [4].

In terms of semantics, adverse events are compatible with signs and symptoms. When a dictionary-based method is used to extract such instances, a binary classifier is needed to differentiate between the signs and symptoms that correspond to adverse events and those associated with the underlying diagnosis [5]. Along similar lines, when an adverse event is associated with medication, a system is needed to support safety evaluators in identifying reports that may demonstrate causal relationships with the suspect medications. To this end, it has been shown that a binary classifier can be trained to successfully differentiate between 2 causality categories: certain, probable, or possible versus unlikely or unassessable [6]. Multifaceted classification can be performed to identify additional properties of an adverse event, for example, temporal (historical or present), categorical (assertive, hypothetical, retrospective, or a general discussion), and contextual (deduced or explicitly stated) [7].

Alternatively, the problem of identifying adverse events can be defined as that of information extraction [8]. More specifically, we can differentiate between entity and relationship extraction.

Here, the goal of entity extraction is to identify a text sequence that describes an adverse event. Therefore, it can also be viewed as a sequence labeling problem [9-11]. In addition, the text sequence can be mapped to a relevant dictionary such as the Medical Dictionary for Regulatory Activities [12,13] or the Unified Medical Language System (UMLS) [9,14]. Such normalization of named entities to standardized identifiers is especially relevant when processing text originating from social media, whose language tends to be highly colloquial [4,9,10,13-17].

When multiple medicines are considered, 2 types of named entities need to be extracted—medicines and adverse events—and additional reasoning needs to be performed to extract a relationship between the two [7,17,18]. Further statistical analysis can be applied to such pairs to measure the strength of such associations [18]. Information of interest can be extracted using pattern-matching approaches, where patterns are typically modeled using regular expressions [7,12,19]. Alternatively, frequent patterns of language for expressing opinions about medications can be learned automatically using association rule mining by considering sentences as transactions and the words in a sentence as items in the transactions [15].

Specific methods chosen to mine adverse events from text depend on the way the text mining problem is posed. Typical approaches chosen for text classification include rule-based methods [3,7,14,20] and supervised machine learning [3-6,16,21]. A range of machine learning methods has been used, including naive Bayes, support vector machines, random forests, maximum entropy, and logistic regression. On occasion, ensemble learning has been used to improve classification performance by integrating multiple models using methods such as bagging, majority voting, weighted averaging, and stacked generalization [4,17,21]. The different types of lexical, syntactic, and semantic features have been used by the classification algorithms. Lexical features include n-grams [4,16], context windows [17], and lexicon matches [16]. Typically, syntactic features include part-of-speech tags, negation, syntactic dependencies, and syntactic functions [16,17,21]. Semantic features are either based on external sources such as the UMLS, PubChem, or DrugBank [16,17,20,22] or manually engineered [4-7]. Other used features were based on sentiment polarities [4,16] and topic modeling [16]. A few examples of using feature selection methods include binormal separation [4] and information gain [17].

Finally, approaches chosen to address adverse event mining as a sequence labeling problem include conditional random fields (CRFs) [9,23] and, more recently, neural networks (NNs) [21,22], including recurrent NNs [10] and long short-term memory (LSTM) [24], which outperformed CRFs. For best results, bidirectional LSTM is combined with CRF [11,25-29]. Most approaches used word embeddings, which represent words as meaningful real-valued vectors of configurable dimensions learned automatically from a large corpus based on their co-occurrence using methods such as word2vec [22,27], fastText [24], and GloVe [30]). Traditional bag-of-words (BOW) approaches tend to struggle with unseen or rare words. Word embeddings that are pretrained on a large corpus remedy this problem and, consequently, boost recall (R).

The aforementioned word-embedding models generate a single embedding for each word, thus conflating homonyms in the corresponding vector space. Bidirectional Encoder Representations from Transformers (BERT) [31] captures contextual relationships in a bidirectional way to contextualize the embedding of any given word based on the surrounding words. BERT is based on an encoder–decoder NN architecture, which can not only be used to generate word embeddings but can also be fine-tuned and further trained for various text mining tasks. For example, it has been used to model adverse event extraction as a named entity recognition (NER) task [11,32]. The topics of word embedding and BERT, in particular, will be revisited later in this paper in the context of motivating and describing our own approach to this problem.

The after-the-fact nature of text data collected from sources such as spontaneous reporting systems, medical literature, electronic health records, and social media naturally gives rise to postmarketing surveillance applications [2,33]. However, pharmacovigilance starts by collecting safety information derived from randomized controlled trials. Our review of text mining applications related to the identification of adverse events revealed that this source of data was underrepresented. This study addresses this gap by using SAE report forms collected during clinical trials as the primary source of data. Given that each trial focuses on a specific medicinal product, the problem is somewhat simplified as the need to extract information about the product itself is obviated. This also makes it more natural to define it as a multi-label text classification problem rather than an information extraction problem. Using the UMLS as our classification scheme, the main aim is to map each document to a set of coded adverse events. The main difficulty of the problem lies in differentiating between signs and symptoms associated with the underlying condition and those that represent adverse events. The fact that both types of references to signs and symptoms can be found within a single SAE report, often within the same sentence, renders a BOW approach unsuitable. Instead, we opt for a deep learning approach. Instead of LSTM approaches, which seem to dominate in our review of the related work, we opt for transformers, which tend to outperform recurrent NNs on a variety of natural language processing tasks.

Methods

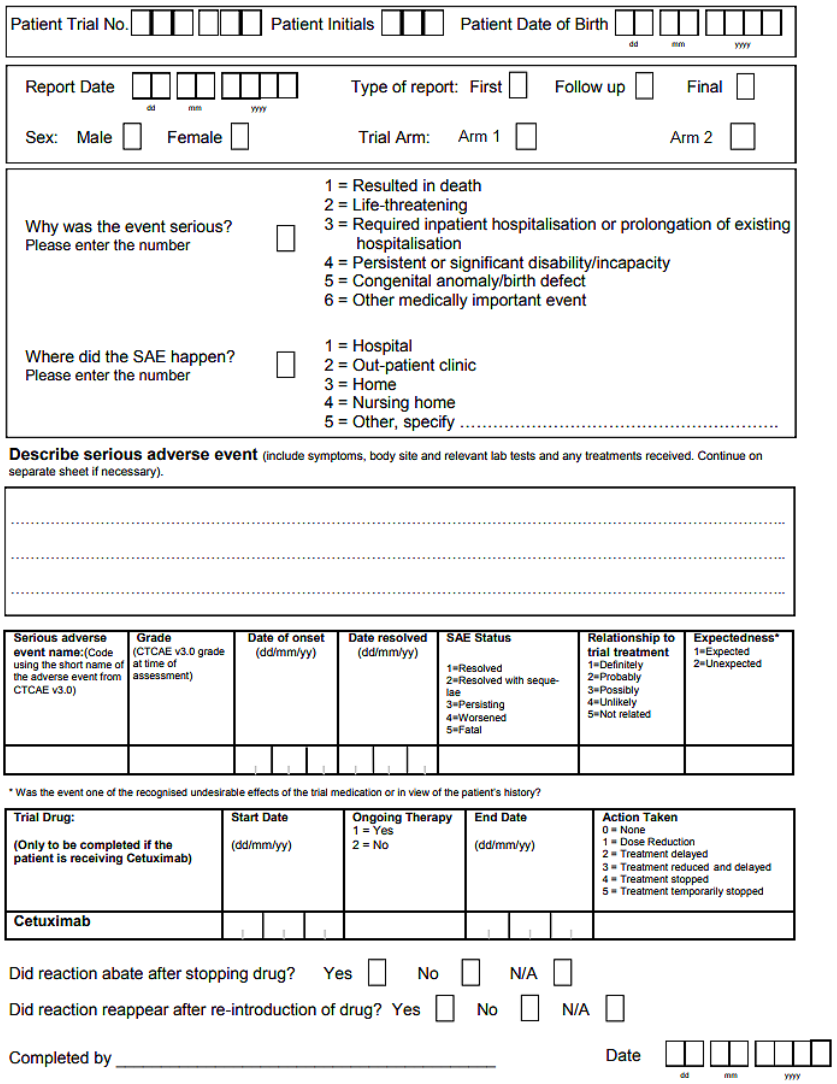
Data Provenance

Data were provided by the Center for Trials Research (CTR), the largest group of academic (noncommercial) clinical trial staff in Wales. Their portfolio of work includes drug trials and complex interventions, mechanisms of disease and treatments, cohort studies, and informing policy and practice in partnerships with researchers across the United Kingdom and worldwide. Across all these trials, standard procedures are put in place to monitor and manage safety reporting and SAE in line with the regulatory requirements for research.

Clinical trials SAE report forms (Figure 1) are completed by research nurses and physicians at hospital or clinical trial sites and submitted as PDF documents to the CTR central safety team for management and processing. They contain data on the

SAE and a narrative description of the event. The narrative is used by the reviewer to help assess causal relationships with the trial drug but is not entered into the trial database and is not used in any analysis of the events. Completed SAE reports are then sent for review by a physician and, depending on the outcome of the review, are logged in the safety databases for the regulatory authorities, ethics committees, and drug companies.

Figure 1. A serious adverse event (SAE) reporting form. CTCAE: Common Terminology Criteria for Adverse Events; N/A: Not Applicable.


Patient Trial No. Patient Initials Patient Date of Birth (dd mm yyyy)

Report Date (dd mm yyyy) Type of report: First Follow up Final

Sex: Male Female Trial Arm: Arm 1 Arm 2

Why was the event serious? Please enter the number
1 = Resulted in death
2 = Life-threatening
3 = Required inpatient hospitalisation or prolongation of existing hospitalisation
4 = Persistent or significant disability/incapacity
5 = Congenital anomaly/birth defect
6 = Other medically important event

Where did the SAE happen? Please enter the number
1 = Hospital
2 = Out-patient clinic
3 = Home
4 = Nursing home
5 = Other, specify .....

Describe serious adverse event (include symptoms, body site and relevant lab tests and any treatments received. Continue on separate sheet if necessary).

Serious adverse event name:(Code using the short name of the adverse event from CTCAE v3.0)	Grade (CTCAE v3.0 grade at time of assessment)	Date of onset (dd/mm/yy)	Date resolved (dd/mm/yy)	SAE Status 1=Resolved 2=Resolved with sequelae 3=Persisting 4=Worsened 5=Fatal	Relationship to trial treatment 1=Definitely 2=Probably 3=Possibly 4=Unlikely 5=Not related	Expectedness* 1=Expected 2=Unexpected

* Was the event one of the recognised undesirable effects of the trial medication or in view of the patient's history?

Trial Drug: (Only to be completed if the patient is receiving Cetuximab)	Start Date (dd/mm/yy)	Ongoing Therapy 1 = Yes 2 = No	End Date (dd/mm/yy)	Action Taken 0 = None 1 = Dose Reduction 2 = Treatment delayed 3 = Treatment reduced and delayed 4 = Treatment stopped 5 = Treatment temporarily stopped
Cetuximab				

Did reaction abate after stopping drug? Yes No N/A

Did reaction reappear after re-introduction of drug? Yes No N/A

Completed by _____ Date (dd mm yyyy)

Although narratives in noncommercial settings, such as CTR, can be digitized, this does not currently take place at the point of initial SAE reporting, as electronic data capture for the SAE report is associated with additional regulatory challenges, primarily because of the requirement for signature verification by a physician and a contemporaneous changelog. Clinical trial staff reviewing SAE reports are, thus, unable to systematically analyze the information provided in the narrative, missing an opportunity to identify the trends and potential safety signals. If the text mining approach were to identify additional safety events and signals not detected through standard reporting, processes could be altered to improve work practices at the level of a noncommercial CTU pharmacovigilance team.

This study aims to assess the feasibility of text mining in the context of such an analysis. The findings could affect the way regulatory narratives are reviewed and analyzed, for example, noncompliances or audit findings.

Data Collection

Data were collected from 6 ongoing clinical trials, as described in Table 1.

Ethical review and approval were waived for this study as this study involved the use of secondary SAE data that were fully deidentified. All involved trials were conducted according to the guidelines of the Declaration of Helsinki and approved by the relevant research ethics committees. All chief investigators from these trials were consulted, and sponsor agreement was obtained for the use of the data in this secondary research study. Participant consent was also waived for the reasons stated above.

A subset of SAE reports was sampled randomly from each trial, giving a total of 286 reports. Phases 1 and 2 were early phases with a smaller number of participants and were not powered. The fewer numbers of reported SAEs were a function of the smaller numbers of participants compared with phase 3; hence,

there were variations in the number of documents across the 6 trials.

The original SAE reports were pseudoanonymized at the point of extraction from the system by obscuring any links between the patient and their individual records. The narrative sections of the SAE reports were then transcribed and saved as Microsoft Word documents. The transcription process was extended to include deidentification by obscuring any personally identifiable information in a way that minimizes the risk of unintended disclosure of the identity of individuals and information about them. The transcribed documents were an average of 37 (SD 24) tokens long.

Table 1. Clinical trials from which data were collected.

ID	Description	Documents, n
Trial-1	A phase 2 study of neoadjuvant chemotherapy given before short-course preoperative radiotherapy as treatment for patients with MRI[a]-staged operable rectal cancer at high risk of metastatic relapse	5
Trial-2	A phase 1b/2 randomized placebo-controlled trial in postmenopausal women with advanced breast cancer previously treated with drug A	7
Trial-3	A randomized phase 3 clinical trial investigating the effect of drug B added to standard therapy in patients with lung cancer	131
Trial-4	Study of chemoradiotherapy in esophageal cancer, plus or minus drug C	34
Trial-5	A phase 1/2 single-arm trial to evaluate combination drugs for the treatment of advanced cancers, including first-line treatment of patients with advanced transitional cell carcinoma of the urothelium	3
Trial-6	A randomized phase 3, open-label, multicenter, parallel group clinical trial to evaluate and compare the efficacy, safety profile, and tolerability of oral drug X versus intravenous drug Y in the treatment of patients with breast cancer and bone metastases	106

[a]MRI: magnetic resonance imaging.

Data Annotation

The aim of this task was to annotate adverse events in the transcribed versions of the SAE report forms. For the purpose of this task, an adverse event was defined as any unfavorable or unintended disease, sign, or symptom (including an abnormal laboratory finding) that is temporally associated with the use of a medical treatment or procedure, which may or may not be considered related to the medical treatment or procedure. Such an event could be related to the intervention, dose, route of administration, or patient or caused by an interaction with another drug or procedure.

The annotation guidelines prescribed the scope of the annotation task as follows: (1) focus only on adverse events that have occurred in the present or past, that is, ignore hypothetical or future events; (2) annotate the entire phrase that describes an adverse event; and (3) if the same adverse event were mentioned multiple times, then annotate every mention. The annotation process was based on the following instructions: (1) identify an adverse event that is mentioned in the narrative, (2) select the text that describes the adverse event, and (3) highlight the selected text.

The text editing operations were performed using Microsoft Word, which was preferred over a specifically designed annotation tool such as BRAT or Bionotate [34] because of zero installation and training overhead. Microsoft Word supports the bulk selection of text based on its formatting. This functionality was used to export highlighted text as standoff annotations, which were later used to calculate the interannotator agreement.

A total of 2 annotators independently annotated all the documents. Figure 2 provides an example. Here, both annotators annotated 2 mentions of tremor but did not annotate the historical mention of tremor as it was not temporally associated with the use of the medical treatment that was the subject of the given clinical trial. Further, 1 reviewer failed to annotate vomiting, leading to disagreement, which was later resolved through discussion. To identify all such cases, we compared all annotations automatically and measured the interannotator agreement.

Figure 2. A serious adverse event report annotated independently by 2 annotators. The annotations are highlighted in yellow.

Annotator A	Annotator B
Post oxaliplatin dose patient began to tremor – patient has a history of tremor when feeling nervous. Vomited x 1 . Admitted for observation as patient lives alone. Tremor now resolved.	Post oxaliplatin dose patient began to tremor – patient has a history of tremor when feeling nervous. Vomited x 1 . Admitted for observation as patient lives alone. Tremor now resolved.

The 2 annotators labeled SAEs as phrases, which were sequences of words whose total number, together with their start and end positions, were not prefixed. Comparing the interannotator agreement at the token level, as suggested by

Tomanek et al [35], was not entirely appropriate for 2 reasons. First, the annotators labeled phrases as sequences of tokens instead of labeling the tokens individually. Therefore, such an approach approximated the original annotation task. More importantly, the number of negative cases (ie, the tokens that had not been annotated) would inevitably be much larger than the number of positive cases, thus skewing the data. The lack of a well-defined number of negative cases prevented the use of traditional interannotator agreement measures such as Cohen κ statistic [36]. A common way of quantifying interannotator agreement in such circumstances is to use information retrieval performance measures instead [37]. By treating one annotator's annotations as the gold standard and the other one's as predictions, we calculated the numbers of true positives (TPs), false positives (FPs), and true negatives, as shown in the confusion matrix (Table 2). When these values were combined to calculate the F1 score, it no longer mattered which annotator was considered the gold standard as this measure was symmetrical.

Table 2. Agreement between 2 annotators.

Positive or negative	Gold positive	Gold negative
Predicted positive	TP[a]=744	FP[b]=50
Predicted negative	FN[c]=98	N/A[d]

[a]TP: true positive.

[b]FP: false positive.

[c]FN: false negative.

[d]N/A: not applicable.

These values can then be used to calculate the precision (P), R, and F1-score as follows (where FN denotes false negative):

$$P=TP/(TP+FP)=744/(744+50)=0.9370$$

$$R=TP/(TP+FN)=744/(744+98)=0.8836$$

$$F1=(2\times P\times R)/(P+R)=0.9095$$

An advantage of using information retrieval performance measures to estimate interannotator agreement is that their values can later be used to gauge a system against human-like performance. At F1=0.9095, the interannotator agreement was found to be relatively high. A total of 148 disagreements were resolved through discussions to establish the ground truth. As part of the discussions, the agreed annotations of adverse events were coded manually against the UMLS, which integrates multiple terminologies, classifications, and coding standards in an attempt to support the interoperability between biomedical information systems, including electronic health records [38]. The MetaThesaurus Browser, a web-based search interface, was used to query the UMLS for each annotation to identify the corresponding concept (Figure 3). This searching procedure involved checking concept definitions to make sure that the chosen concept matched the sense of the adverse event annotation. Each concept in the UMLS is assigned a concept unique identifier (CUI), which was used to code the corresponding annotation (see Figure 4 for examples). Subsequently, the CUI codes were extracted, duplicates were removed, and the remaining CUIs were used as class labels for each document. Table 3 provides a statistical summary of the annotated data set, which contains a total of 995 class labels.

Figure 3. Metathesaurus browser search results.

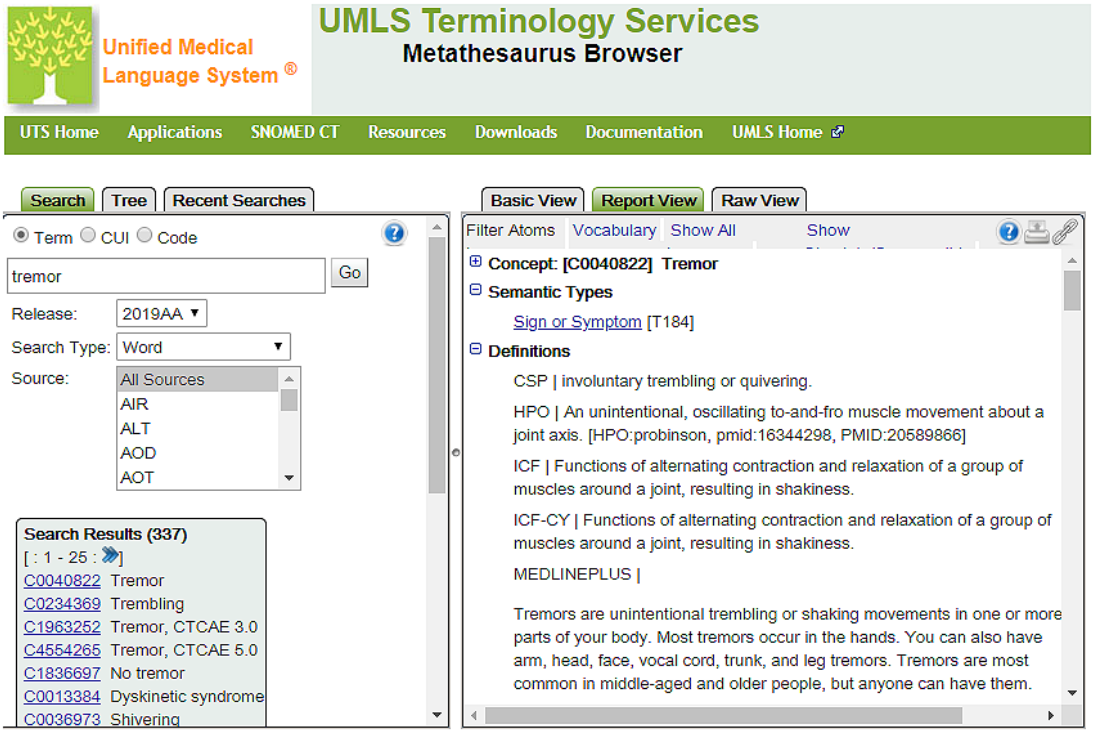


Figure 4. Coding of documents against the Unified Medical Language System.

Document	Labels
Post oxaliplatin dose patient began to tremor C0040822 – patient has a history of tremor when feeling nervous. Vomited C0042963 x 1. Admitted for observation as patient lives alone. Tremor C0040822 now resolved.	C0040822 C0042963

Table 3. Statistical properties of the annotated data set.

Statistical properties	Document length (in tokens)	Annotations	Class labels
Values, minimum	2	1	1
Values, maximum	223	20	19
Values, median	31	3	3
Values, mean (SD)	36.71 (23.77)	3.76 (2.46)	3.48 (2.18)

Problem Representation

The aim of this study was to automate the identification of adverse events described in the narrative section of the SAE reports. This goal was cast as a text classification problem. Given a document and classification scheme, the system should label the document with the relevant classes from the given scheme. In our case, the document was an SAE report, a classification scheme was the set of concepts encompassed by the UMLS, and their CUIs were used as class labels. The second column in Figure 4 provides an example of the expected output.

To identify the possible adverse events mentioned in a document, the first step involved looking for concepts of the relevant semantic types. In our approach, the UMLS dictionary lookup was restricted to 6 manually selected semantic types: disease or syndrome, finding, injury or poisoning, neoplastic process, pathological function, and sign or symptom. Some of their mentions could be in the context of medical history and,

therefore, not necessarily constitute an adverse event. To differentiate between the 2 types of mentions, we formulated a binary classification task at the concept level: given a context, does a specific UMLS concept constitute an adverse event? Figure 5 provides different references to the concept of *pleural effusion*. For example, the first 3 references do not constitute adverse events. The first and third mentions of *pleural effusion* refer to medical history, whereas the second mention is negated. The remaining 3 mentions of *pleural effusion* refer to the cause of hospital admissions that prompted SAE reporting.

The practical implementation of such problem representations started with linguistic preprocessing, which was originally developed to support cohort selection from hospital discharge summaries, adapted for this study [39]. This module involved text segmentation and basic string operations such as lowercasing, fully expanding enclitics and special characters, replacing a selected subset of words and phrases with their representatives, and, in particular, replacing acronyms and abbreviations with their full forms. Finally, the preprocessed documents were analyzed using MetaMap [40], a highly configurable dictionary lookup software, to find mentions of UMLS concepts from the 6 semantic types listed above. Figure 6 illustrates a portion of the UMLS dictionary and how it was matched against the input text. As the figure illustrates, a single document might contain multiple adverse events. To support the classification of one adverse event candidate at a time, a separate copy of the given document was saved for each candidate. Each copy anchored a single concept, which may have had multiple occurrences, by marking them up in line. In addition, the text was further regularized by replacing all the concepts with their preferred names. Concept anchoring provided a simple, uniform representation of the potential adverse events, which enabled us to train a single binary classifier based on the context surrounding the anchors.

Figure 5. Adverse event identification as a binary classification task. CT: computed tomography.

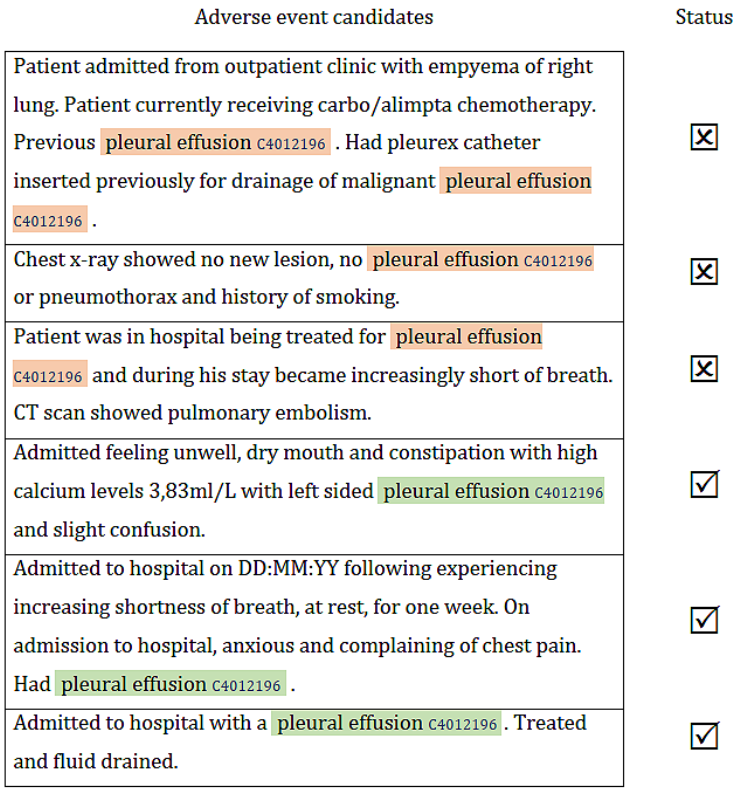
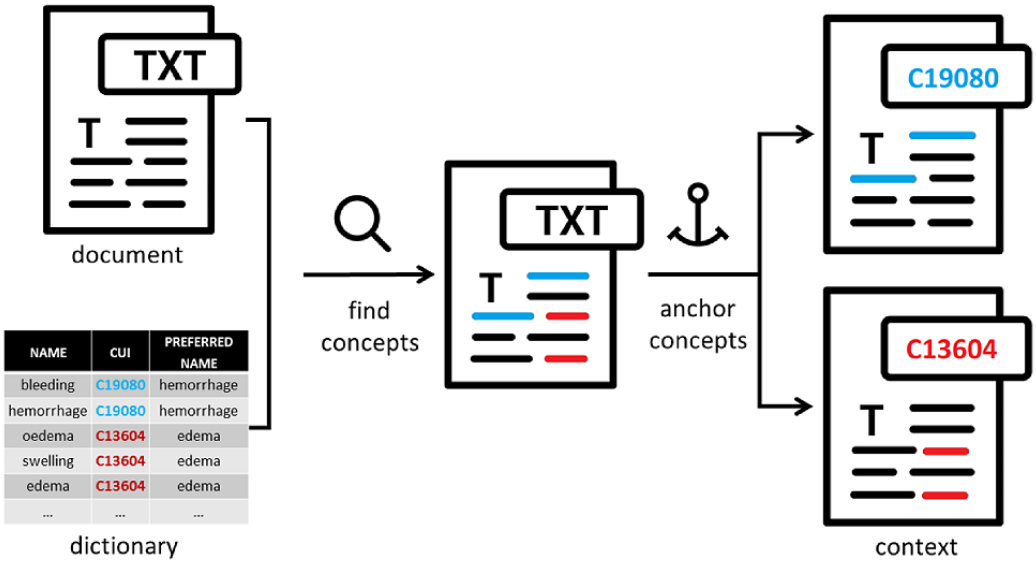


Figure 6. Identification of potential adverse event mentions. CUI: concept unique identifier.



Classification Rationale

The binary task formulation itself—*given a context, does a specific UMLS concept constitute an adverse event?*—indicates 2 main types of involved features: extrinsic (context) and intrinsic (concept). Extrinsic features may include the number of mentions within a document, the position within a document, and other words within a fixed-size window. When combined with gold standard annotations, machine learning can be used to discover how to differentiate between positive and negative contexts without having to manually describe the patterns of positive and negative use. For example, by considering the co-occurring words (see Figure 7 for examples) and the corresponding annotations, a simple NN can learn to use words such as *previous* and *have* as negative and positive modifiers, respectively. By considering a wider context, more complex patterns such as *admitted to hospital with* and *known to have* (see Figure 8 for examples) would start to emerge as positive and negative contexts, respectively. Traditionally, such patterns were observed using corpus linguistics methods, which were engineered manually and encoded formally as regular expressions [41]. In recent times, NNs are used to automatically capture both short- and long-range dependencies.

Similarly, lexical morphology could be explored in an NN approach to learn the patterns of subwords within a concept's name, which were positively or negatively correlated with adverse events. For example, it is reasonable to expect that any concept identified as a potential adverse event that contains the word *chronic* (eg, *chronic obstructive airway disease* or *chronic infection*) is more likely to refer to a process than a single event. Similarly, any concept whose name contains a word *loss* (eg, *loss of appetite* or *hair loss*) is more likely to be an adverse event. The words themselves can be analyzed for affixes. For example, the prefix hypo- (low or below normal) can be used to increase the likelihood of concepts such as *hypocalcemia* or *orthostatic hypotension* corresponding to adverse events. Similarly, the suffix *-emia* (presence in the blood) can be used to identify concepts such as *cerebrovascular ischemia* or *hyperkalemia* as strong candidates for adverse events. Again, no prior medical knowledge is required to embed such features into NNs, which consider inputs and outputs simultaneously to support end-to-end learning and, hence, bypasses manual feature engineering.

Figure 7. Observing the patterns of positive and negative modifiers. CRTI: common respiratory tract infection; GI: gastrointestinal; OGD: oesophagogastroduodenoscopy; PR: per rectum; SAE: serious adverse event.

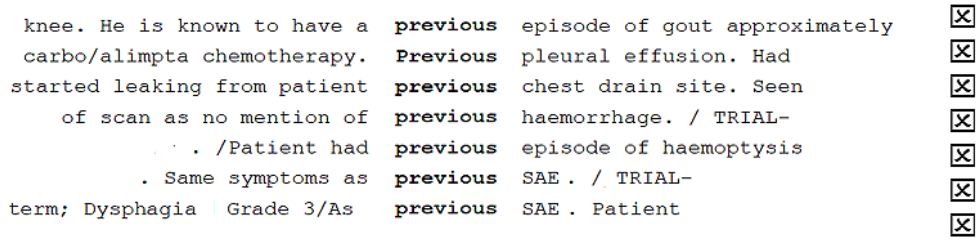
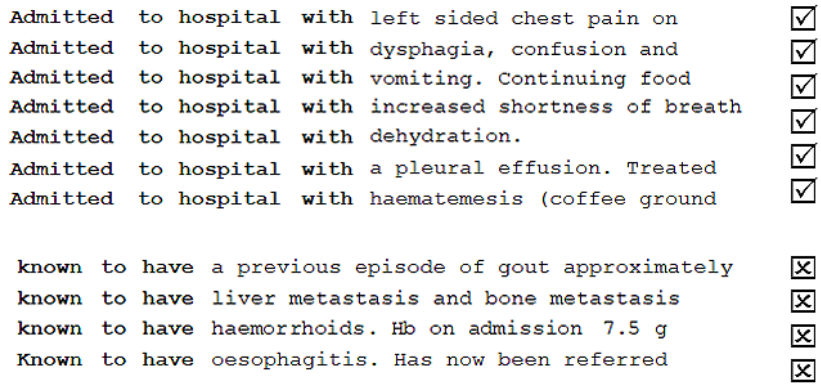


Figure 8. Observing more complex patterns of positive and negative use. Hb: hemoglobin.



Text Representation

The first choice en route to implementing a binary adverse event classifier is text representation. Traditionally, the BOW representation, which is based on the frequency of occurrence of individual words, has been used to support text classification. Given that multiple signs and symptoms, some of which can be adverse events, are commonly discussed in an SAE report, the BOW representation would make it difficult to distinguish adverse events from other signs and symptoms discussed within the same document as it does not preserve local context. In addition, the BOW representation is not robust with respect to the out-of-dictionary problem; that is, any classifier trained using this representation will not be able to use words that were previously not encountered in the training data.

Word embedding can alleviate this problem. Word embedding is a mapping from the lexicosemantic space of words to the n-dimensional real-valued vector space. Methods such as word2vec [42] and GloVe [43] for learning word embeddings from large corpora rely on the hypothesis of distributional semantics, which claims that words occurring in similar contexts tend to convey similar meanings [44]. In other words, these methods assume that the meaning of a word depends on its context, that is, the frequency of co-occurrence with other words within a text window. Consequently, word embeddings tend to arrange semantically related words in similar spatial patterns. Therefore, by mapping a word to its embedding, it becomes possible to model its semantics numerically and thus use arithmetic operations to reason about it. This property is effectively used by NNs in which text is passed through a series of layers that each combines and transforms embeddings to eventually derive an output such as a class label in text classification or an answer in question answering.

Context-free word-embedding models such as word2vec [42] and GloVe [43] generate a single embedding for each word, making it impossible to differentiate between homonyms in the corresponding vector space. For example, the word *mole* would have a single embedding regardless of its many different meanings. Context-sensitive word-embedding models such as BERT [31] generate an embedding for each word based on the surrounding words. For example, the word *mole* used as *a unit of measurement* and *a disorder that affects the soft tissue* will have different representations in the word-embedding space.

BERT [31] is a transformer-based language model that captures contextual relationships in a bidirectional way. A transformer

[45] is an encoder–decoder NN architecture that uses attention mechanisms to forward a holistic interpretation of a sequence to the decoder simultaneously rather than sequentially, as is the case in recurrent NNs such as LSTM and gated recurrent units. For each word, which is represented by its embedding, the self-attention layer considers other words, including their positions, in the same sentence to improve its encoding. As a workaround for the self-attention issue, BERT uses masked language modeling, that is, hides a certain percentage of the words using a special token [MASK] and uses their position to infer these words. The context-sensitive nature of BERT embeddings makes this language model perfectly suited for practical implementation of the classification rationale described earlier. In addition, BERT uses WordPiece tokenization to obtain subword units by applying a greedy segmentation algorithm to minimize the number of WordPieces in the training corpus [46]. This implies that the downstream classification model may be able to use the word morphology.

Classification Model

The masked language modeling was 1 of the 2 tasks on which BERT was trained simultaneously. The second task was the next sentence prediction. In addition to [MASK], BERT uses 2 other special tokens for fine-tuning and specific task training: (1) a classification token [CLS], which indicates the beginning of a sequence and is commonly used for classification tasks (the output associated with this token is used for the next sentence prediction task); and (2) a sequence delimiter token [SEP], which indicates the end of a segment.

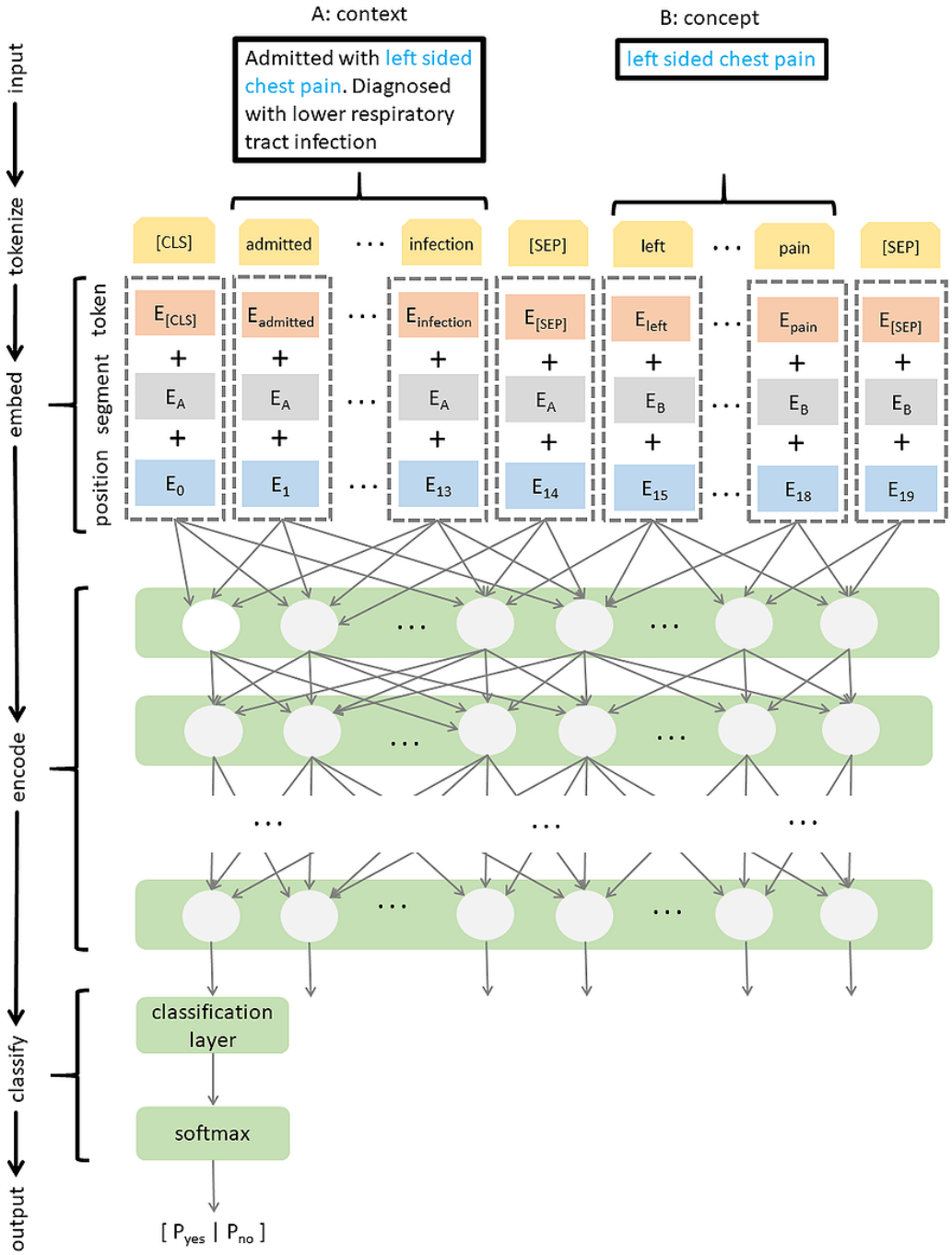
The embedding layer shown in Figure 9 illustrates the input format that BERT expects. Each token's vocabulary identifier is mapped to a token embedding that is learned during training. Next, a binary vector is used to differentiate between 2 text segments, typically sentences. The type of segment depends on a specific task, for example, in question answering both question, and the reference text could be appended and separated by a special delimiter token [SEP]. In our model, we chose the anchored concept as one segment and its context (ie, the whole document) as another. The binary vector was mapped to a segment embedding using a lookup table, which was learned during training. Finally, local token positions were mapped to positional embeddings using a lookup table, which was updated during training.

The 3 types of embeddings were added and fed into the pretrained $BERT_{BASE}$ model, which comprises 12 layers of transformer encoders, each having a hidden size of 768 and 12 attention heads. Each layer produces a token-specific output, which can be used as its (contextualized) embedding. Similar to binary classification tasks described in [31], the final transformer output corresponding to the special [CLS] token was taken as an aggregate problem representation, that is, pooled output, and passed on to the classification layer after a 0.1 dropout, which was used to reduce overfitting.

The classification layer reduced the size of the pooled output from 768 to 2, which corresponds to the log-odds (or logits) of the classification output with respect to the question of whether the given concept was an adverse event or not. In contrast to the network up to that point, the classification layer was not pretrained. Instead, the corresponding weights were learned during BERT fine-tuning. As suggested in the study by Devlin [31], the weights were initialized using a truncated normal distribution with mean 0 (SD 0.02). A softmax function was then applied to obtain the probability distribution of the 2 classes. The loss function (softmax cross entropy between the logits and the class labels) was optimized using the Adam optimizer with an initial learning rate of 2×10^{-5}, which was chosen without any fine-tuning, based on the values suggested in the study by Devlin [31].

The classification model was trained for 8 epochs. This hyperparameter was preselected without any tuning. In each epoch, the training data were looped over in batches of 8 samples. The batch size was limited by memory. All other parameters were kept identical to those in the original $BERT_{BASE}$ uncased model, including the clip norm of 1.0, and linear warmup (100 warmup steps with linear decay of learning rate). The system was implemented in TensorFlow [47], an open-source software library for machine learning, with a particular focus on training and inference of deep NNs, using the GeForce RTX 2080 (Nvidia Corp) graphics processing unit to accelerate deep learning.

Figure 9. Architecture based on Bidirectional Encoder Representations from Transformer (BERT) for classification of adverse events. CLS: classification token; SEP: sequence delimiter token.



Results

During preprocessing, MetaMap was used to extract adverse event candidates. MetaMap failed to extract a total of 118 adverse events from the ground truth. Therefore, these instances automatically constituted FNs. The remaining 1021 adverse event candidates extracted by MetaMap were passed on to the BERT-based classification model shown in Figure 9. To understand the performance of the BERT classifier, we first focused only on these 995 adverse event candidates before amalgamating them with 118 FNs. Of the 995 candidates, 659 (66.2%) were positive instances (ie, regarded as adverse events in the ground truth), and 336 (33.8%) were negative instances (ie, not regarded as adverse events in the ground truth).

We performed 10 independent 5-fold cross-validations to evaluate the performance of the classification model. In other words, during each cross-validation, 20% of the documents were held out for evaluation, whereas the remaining 80% were used for training, and this was done 5 times in a row, each time using a different fold for evaluation. More specifically, for each of the 10 independent runs, we did the following:

The 286 unique document identifiers were first shuffled randomly and then split into 5 folds. Remember that each document may have contained multiple adverse event

candidates, and a separate copy was created for each candidate during preprocessing. All copies of the same document shared the same document identifier; hence, there was no overlap of data across the folds. As the splitting was done by document irrespective of the number of events they contained, the actual number of samples (ie, potential adverse events identified by MetaMap) in each fold may vary. We looped over the folds, each time using a different fold for evaluation and the remaining 4 folds for training. Each time, we measured P, R, and F1 scores. Once each of the 5 folds was used for evaluation, we calculated the mean values obtained for each evaluation measure. Finally, these values were averaged over 10 independent runs.

The same cross-validation process was applied to the baseline approach. Remember that the goal of our system was to code adverse events against the UMLS; therefore, a UMLS lookup was inevitable. The lookup itself could be performed as the first step to identify an adverse event candidate (and code it at the same time) and then classify it. Alternatively, it could be performed as the last step to code an adverse event, which was first extracted from free text. In the former approach, we were dealing with a binary classification problem where it needed to be determined whether a given UMLS concept was an adverse event or not. In the latter approach, we were dealing with a sequence labeling problem where the boundaries of a token sequence that referred to an adverse event needed to be determined. This is how Du et al [32] approached the extraction of adverse events from safety reports by framing it as the NER problem and fine-tuning BERT for this task. We reimplemented and cross-validated their approach on our data set to establish the baseline. Although the authors originally used BERT for biomedical text mining (BioBERT) [48], we replaced it with BERT in our experiments to make their approach directly comparable with ours. The results achieved by the 2 contrasting approaches are presented in Table 4. Despite the similarities in the underlying technologies, we can observe a notable difference in the performance of the 2 approaches, most prominently in terms of P, where we can see an improvement of approximately 30 percent points over the baseline. A detailed analysis of this phenomenon is provided in the *Discussion* section. In this section, we proceed to describe the results achieved using our own approach.

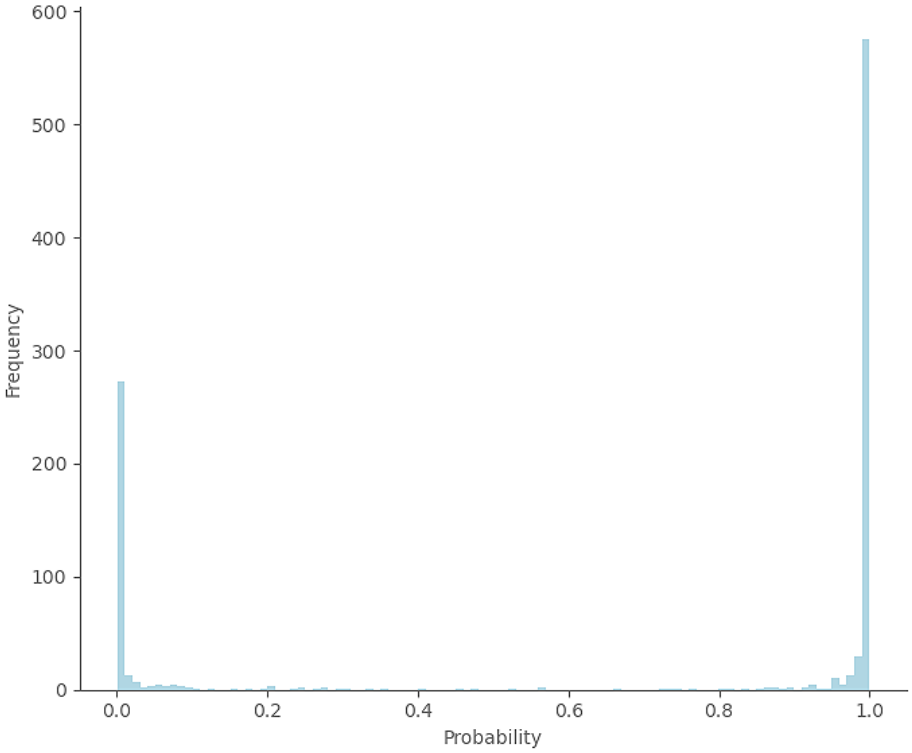
Figure 10 displays the distribution of the prediction probabilities. The histogram combines the predictions from all folds used for cross-validation. We can observe that most prediction probabilities are concentrated around the 2 extremes, 0 and 1, which suggests that the classification model is able to make clear-cut decisions, as it does not depend on a specific threshold.

Table 4. Evaluation results.

Parameters	Baseline approach: named entity recognition (BERT[a])+concept extraction (MetaMap), mean (SD)	Our approach: concept extraction (MetaMap)+classification (BERT), mean (SD)
Precision	0.5715 (0.0076)	0.8638 (0.0057)
Recall	0.7116 (0.0096)	0.7604 (0.0121)
F1 score	0.6335 (0.0072)	0.8080 (0.0071)

[a]BERT: Bidirectional Encoder Representations from Transformers.

Figure 10. Distribution of prediction probabilities for all folds in a cross-validation experiment.



In Figure 11, we used receiver operating characteristic curves to illustrate the diagnostic ability of the classification model. A separate curve was provided for each of the 5 folds used for cross-validation. The plot shows the TP rate versus the FP rate at each classification threshold. The solid-colored lines correspond to the model's performance, whereas the gray dashed line represents the performance of a classifier with no skill, that is, the one that always predicts the majority class. An ideal model would result in a curve that bows toward the coordinate (1,0). With its curve consistently lying close to the top-left corner, our model demonstrated very good classification performance. We summarized the receiver operating characteristic results by calculating the area under the curve to measure the ability of our model to distinguish between the 2 classes, with higher values indicating better performance. With an overall mean score of 0.8789 (SD 0.0101) and a range between 0 and 1, our model was clearly able to distinguish between adverse events and underlying conditions 87.79% of the time on average.

Finally, to account for the class imbalance, we also looked at the precision-recall (PR) curve shown in Figure 12. Again, the solid-colored lines correspond to our model's performance, whereas the gray dashed horizontal line corresponds to a model with no skill, that is, a model whose P is equal to the proportion of positive samples. The PR curve of our model was relatively close to that of an ideal model, whose curve would bow toward the coordinate (1,1). In comparison to a no skill model, which would achieve a PR area under the curve score of 0.6533, our model reached a high score of 0.9108 (SD 0.0103), demonstrating its ability to correctly classify adverse events despite the class imbalance.

Figure 11. Receiver operating characteristic curve for each fold in a cross-validation experiment.

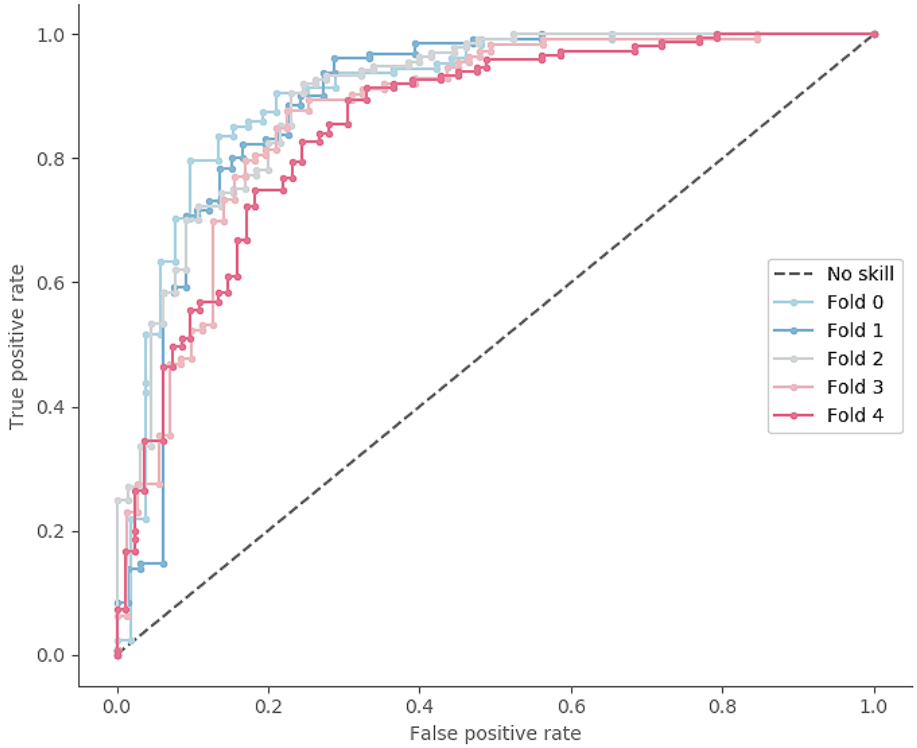
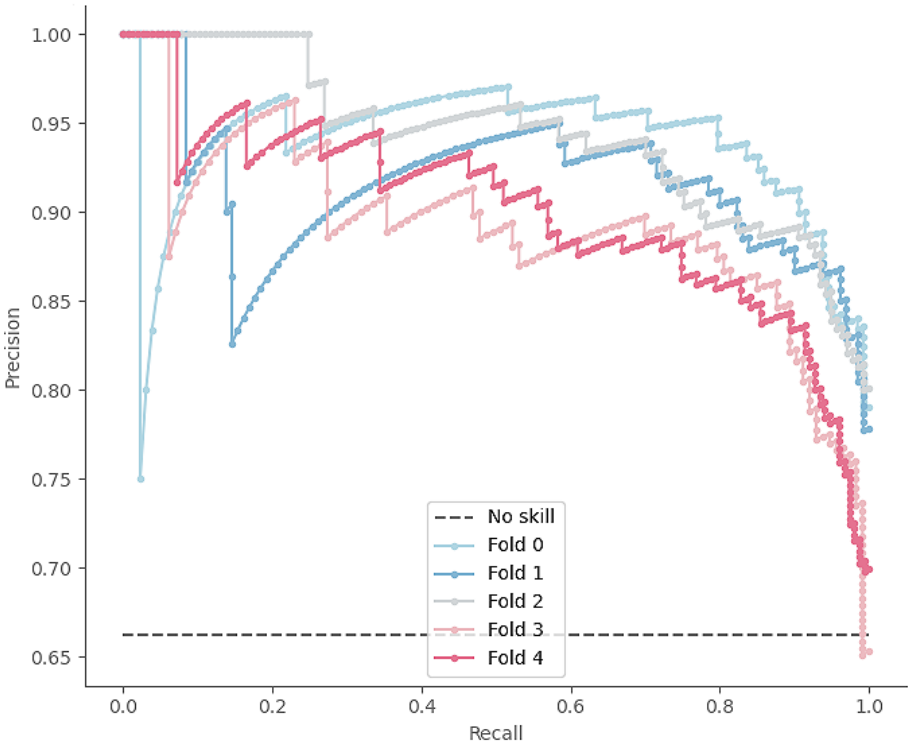


Figure 12. Precision-recall curve for each fold in a cross-validation experiment.



Discussion

Principal Findings

Previously, we provided details on calculating the interannotator agreement using P, R, and F1 score. When a system is evaluated against the ground truth, the corresponding values establish the human performance baseline, which in this case were P=0.9370, R=0.8836, and F1=0.9095. If we compare these values against the results provided in Table 4, we can observe a 10.15 percent points difference in the F1 score. In particular, we notice that the system's R is 10.34 percent points lower than its P. There are 2 potential sources of type 2 errors in the system. Remember that the system first uses MetaMap to identify potential adverse events, which are then classified by BERT as positive or negative. Both components can give rise to FN results. First, any adverse event that MetaMap failed to forward to BERT would have been automatically counted as an FN. Second, any adverse event that MetaMap did supply to BERT for further classification could have still ended in an FN. MetaMap is a predefined rule-based system, and as such, its performance within our system is limited by external factors. BERT, on the other hand, has been trained for a specific task using the data set described here. Therefore, it is worth focusing specifically on its classification performance.

To evaluate how well BERT learned to classify adverse events, we removed those FNs from the ground truth that were never actually classified by BERT because of MetaMap failing to identify them in the first place. Table 5 provides the cross-validation results for BERT's performance alone. We observe that the classification performance alone is much closer to the human performance baseline, lagging behind the F1 score by only 2.93 percent points.

Table 5. Bidirectional Encoder Representations from Transformers' (BERT) performance.

Parameters	Named entity recognition (BERT), mean (SD)	Classification (BERT), mean (SD)
Precision	0.7484 (0.0066)	0.8651 (0.0053)
Recall	0.8237 (0.0086)	0.8974 (0.0104)
F1 score	0.7835 (0.0053)	0.8802 (0.0044)

If we now compare BERT's classification performance given in Table 5 with the overall system performance given in Table 4, we can see that the P is virtually identical (0.8638 vs 0.8651), whereas R differs by 13.70 percent points (0.7604 vs 0.8974). Hence, we can conclude that the R of the overall system is primarily limited by MetaMap's performance, which naturally raises the question of whether its use as a preprocessing step within our system was appropriate. The baseline method uses MetaMap as the postprocessing step; therefore, we investigated the extent of its effect on the overall performance by singling out BERT's performance on the NER task, which was evaluated using the exact matching of phrases annotated in the ground truth. If we compare the first column of Table 5 with the second column of Table 4, we can observe that without MetaMap, BERT can certainly achieve higher R (0.8237 vs 0.7604) when it is allowed to determine the phrase boundaries on its own rather than having them prescribed by MetaMap.

Although such an approach is unarguably more flexible, it can also have a negative impact when the goal of the system is to code adverse events rather than only recognize their mentions in the text. If the phrase boundaries are not correctly detected as part of the NER task, then searching the UMLS using an incorrectly extracted phrase may provide an incorrect code. Consider, for example, 2 adverse events, *respiratory tract infection* (whose code in the UMLS is C0035243) and *urinary tract infection* (whose code is C0042029). Suppose that a system failed to correctly identify their boundaries, for example, by suggesting *tract infection* in both cases. The UMLS has no concept referring to *tract infection*; therefore, MetaMap would at best suggest *infection* (whose code is C3714514) as the closest concept matching the given search term, thus incorrectly coding both *respiratory tract infection* and *urinary tract infection*, resulting in 2 FNs (labeled C0035243 and C0042029 in the ground truth) and 2 FPs (both labeled C3714514 by the system). On the other hand, MetaMap can be configured to recognize the longest phrases from relevant semantic types and, in that way, impose tighter control of the process, reducing the number of both FPs and FNs. Although MetaMap may limit R, it does play an important role in controlling the P in our proposed approach, as the results in Table 4 clearly depict. Nonetheless, MetaMap could benefit from revising its rule-based dictionary lookup approach in light of the new advances in text mining and, in particular, deep learning approaches to bring its performance in line with the state of the art.

Focusing on BERT's performance alone in Table 5, we can see that it performs better on the binary classification task than the NER task. This is not surprising, as the sequence labeling task is inherently more complex than binary classification. This is because of the number of possible sequences growing exponentially with the length of a document. In particular, the performance gap is bound to widen when training the corresponding models on a relatively small data set, as is the case in this study. Having <300 annotated documents available, we can see from Table 5 that BERT's performance on the classification task is in the high 80s across all metrics, whereas its performance on the NER task is in the high 70s overall. This again justifies our choice to run BERT after MetaMap rather than the other way around.

Going back to the BERT's classification performance provided in Table 5, while examining the misclassified examples, we noticed some patterns. Some simple negation patterns were not captured by the classifier. For example, in the document containing the sentence "Chest X-ray showed no new lesion, no pleural effusion disorder or pneumothorax and history of smoking," both *pleural effusion disorder* and *pneumothorax* were misclassified as adverse events. Similarly, in the document with the sentence "admitted with right scaptula/back pain, no

chest pain or dyspnea," both *chest pain* and *dyspnea* were misclassified as adverse events.

This finding is in line with the current evidence that neural models struggle to generalize negation to out-of-sample data sets, even within the same domain [49]. The generalizability of negation remains a challenge, as none of the factors considered, including the annotation guidelines, the amount of data available, and their lexical and syntactic properties, fully explained the poor performance [50]. Empirical evidence suggests that the use of domain-specific embeddings such as BioBERT [48] may improve negation detection [51]. BERT can also be fine-tuned to support the negation detection task in clinical text [51,52]; however, this requires data to be annotated specifically for this task. Nonetheless, manual adaptation, be it rule modification or in-domain data annotation, remains a recommended strategy for optimizing performance in clinical natural language processing [50]. Rule-based systems for negation detection such as ConText [53] seem to transfer well within a domain [54]. Therefore, the simplest and most effective way of addressing negation as the source of errors in our proposed framework would be to use the ConText algorithm [53] to detect negated contexts and automatically exclude them from further consideration.

Some words, such as the word *decreasing*, can have the opposite effect depending on the context in which it is used. For example, *decreased mobility* implies a negative effect, whereas *decreased pain* implies a positive effect and not an adverse event. The system was not able to differentiate between such contexts. This could be remedied by incorporating domain knowledge about candidate adverse events. Alternatively, with a larger training data set, these properties could be learned directly from the data.

Finally, the classification model struggled when a given concept was used in multiple contexts. For example, for the concept *infection* in the document extract "admitted to hospital with lower respiratory tract infection [...] not commenced chemotherapy related infection," the model misinterpreted the latter mention as a negated one and, consequently, misclassified this adverse event.

Conclusions

This study established the feasibility of automated coding of adverse events described in the narrative section of the SAE reports. This, in turn, enables statistical analysis of adverse events and the patterns of such events so that any correlations with the use of medicines can be estimated in a timely fashion. An easy adaptation of an existing deep learning architecture trained on a relatively small data set demonstrates that similar tools can be built rapidly. In addition, the evaluation results show that such tools also perform with high accuracy. This performance can be attributed to the choice of the method. BERT is already pretrained on a large unlabeled corpus, which allows it to be fine-tuned on a small, labeled corpus for a specialized task. This is particularly relevant for clinical text mining applications, where the data annotation bottleneck has been identified as one of the key obstacles to machine learning approaches for clinical text mining [55].

Unfortunately, the relevant data are still mainly handwritten, which means that they cannot be immediately processed in the way proposed in this study. There are 2 ways in which this issue can be addressed. We can work with the stakeholders to change the policy on the means of collecting information on SAEs, for example, by transcribing the notes when they reach the safety and pharmacovigilance teams in the central trial unit, by requiring them to be typed, or by using some combination of these 2 approaches.

Alternatively, we can propose to develop methods to digitize handwritten notes automatically using tools such as Transkribus [56], which have been designed to digitize historical documents and allow the training of specific text recognition models. This would have a great potential for impact on safety by digitizing and mining legacy data from previous trials, where some medicinal products may have already reached the market, thus exposing the population to previously overlooked safety concerns. Currently, these issues prevent a systematic analysis of the information provided in the narrative of SAE reports, hence missing an opportunity to identify potential safety signals.

Acknowledgments

The authors are thankful to Kelly Gee for providing advice and guidance on regulatory expectations and best practices for safeguarding patient safety in clinical trials. This study was funded by the Engineering and Physical Sciences Research Council via the Healthcare Text Analytics Network (HealTex), grant number EP/N027280/1. The Centre for Trials Research receives infrastructure funding from the Health and Care Research Wales and Cancer Research United Kingdom.

Authors' Contributions

This study was conceptualized by IS; its methodology was developed by IS, DC, MST, and PC. The software work was handled by DC, IS, and PC. The validation was performed by NA, CJ, and MB; the investigations were done by IS and MB; resources were collected by IS and MB; and the data were curated by NA and IS. The original draft of this paper was prepared by IS and, it was reviewed and edited by IS, DC, MST, PC, and MB. The visualizations were created by IS and DC. The study was supervised by IS and MST, with project administration performed by CJ and funding acquisition managed by IS and MB. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

References

1. Data Mining at FDA - White Paper. US Food and Drug Administration. 2018. URL: https://www.fda.gov/science-research/data-mining/data-mining-fda-white-paper [accessed 2021-12-11]
2. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy 2018 Aug 22;38(8):822-841. [doi: 10.1002/phar.2151] [Medline: 29884988]
3. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. J Am Med Inform Assoc 2011;18(5):631-638 [FREE Full text] [doi: 10.1136/amiajnl-2010-000022] [Medline: 21709163]
4. Chee BW, Berlin R, Schatz B. Predicting adverse drug events from personal health messages. AMIA Annu Symp Proc 2011;2011:217-226 [FREE Full text] [Medline: 22195073]
5. Botsis T, Buttolph T, Nguyen MD, Winiecki S, Woo EJ, Ball R. Vaccine adverse event text mining system for extracting features from vaccine safety reports. J Am Med Inform Assoc 2012;19(6):1011-1018 [FREE Full text] [doi: 10.1136/amiajnl-2012-000881] [Medline: 22922172]
6. Han L, Ball R, Pamer C, Altman R, Proestel S. Development of an automated assessment tool for MedWatch reports in the FDA adverse event reporting system. J Am Med Inform Assoc 2017 Sep 01;24(5):913-920 [FREE Full text] [doi: 10.1093/jamia/ocx022] [Medline: 28371826]
7. Iqbal E, Mallah R, Rhodes D, Wu H, Romero A, Chang N, et al. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. PLoS One 2017 Nov 9;12(11):e0187121 [FREE Full text] [doi: 10.1371/journal.pone.0187121] [Medline: 29121053]
8. Roberts K, Demner-Fushman D, Tonning JM. Overview of the TAC 2017 adverse reaction extraction from drug labels track. In: Proceedings of the Text Analysis Conference (TAC). 2017 Presented at: Proceedings of the Text Analysis Conference (TAC); Nov 13-14, 2017; Gaithersburg, Maryland, USA URL: https://dblp.org/rec/conf/tac/RobertsDT17.html
9. Nikfarjam A, Sarker A, O'Connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. J Am Med Inform Assoc 2015 May;22(3):671-681 [FREE Full text] [doi: 10.1093/jamia/ocu041] [Medline: 25755127]
10. Cocos A, Fiks A, Masino A. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc 2017 Jul 01;24(4):813-821 [FREE Full text] [doi: 10.1093/jamia/ocw180] [Medline: 28339747]
11. Fan Y, Zhou S, Li Y, Zhang R. Deep learning approaches for extracting adverse events and indications of dietary supplements from clinical text. J Am Med Inform Assoc 2021 Mar 01;28(3):569-577 [FREE Full text] [doi: 10.1093/jamia/ocaa218] [Medline: 33150942]
12. Duke J, Friedlin J. ADESSA: a real-time decision support service for delivery of semantically coded adverse drug event data. AMIA Annu Symp Proc 2010 Nov 13;2010:177-181 [FREE Full text] [Medline: 21346964]
13. Combi C, Zorzi M, Pozzani G, Arzenton E, Moretti U. Normalizing spontaneous reports into MedDRA: some experiments With MagiCoder. IEEE J Biomed Health Inform 2019 Jan;23(1):95-102. [doi: 10.1109/JBHI.2018.2861213] [Medline: 30059326]
14. Emadzadeh E, Sarker A, Nikfarjam A, Gonzalez G. Hybrid semantic analysis for mapping adverse drug reaction mentions in tweets to medical terminology. AMIA Annu Symp Proc 2018 Apr 16;2017:679-688 [FREE Full text] [Medline: 29854133]
15. Nikfarjam A, Gonzalez GH. Pattern mining for extraction of mentions of Adverse Drug Reactions from user comments. AMIA Annu Symp Proc 2011;2011:1019-1026 [FREE Full text] [Medline: 22195162]
16. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform 2015 Feb;53:196-207 [FREE Full text] [doi: 10.1016/j.jbi.2014.11.002] [Medline: 25451103]
17. Liu J, Zhao S, Zhang X. An ensemble method for extracting adverse drug events from social media. Artif Intell Med 2016 Jun;70:62-76. [doi: 10.1016/j.artmed.2016.05.004] [Medline: 27431037]
18. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc 2009;16(3):328-337 [FREE Full text] [doi: 10.1197/jamia.M3028] [Medline: 19261932]
19. Skentzos S, Shubina M, Plutzky J, Turchin A. Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository. AMIA Annu Symp Proc 2011;2011:1270-1279 [FREE Full text] [Medline: 22195188]
20. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. Vaccine 2009 Mar 23;27(14):2077-2083. [doi: 10.1016/j.vaccine.2009.01.105] [Medline: 19428833]
21. Negi K, Pavuri A, Patel L, Jain C. A novel method for drug-adverse event extraction using machine learning. Informatics Med Unlocked 2019;17:100190. [doi: 10.1016/j.imu.2019.100190]
22. Wang C, Lin P, Cheng C, Tai S, Kao Yang Y, Chiang J. Detecting potential adverse drug reactions using a deep neural network model. J Med Internet Res 2019 Feb 06;21(2):e11016 [FREE Full text] [doi: 10.2196/11016] [Medline: 30724742]
23. Tao C, Lee K, Filannino M, Buchan K, Lee K, Arora T. Extracting and normalizing adverse drug reactions from drug labels. Semantic Scholar. URL: https://tinyurl.com/bdetz4dw [accessed 2021-12-11]

24. Cocos A, Masino A. Combining rule-based and neural network systems for extracting adverse reactions from drug labels. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA URL: https://tac.nist.gov/publications/2017/participant.papers/TAC2017.CHOP.proceedings.pdf
25. Belousov M, Milosevic N, Dixon W, Nenadic G. Extracting adverse drug reactions and their context using sequence labelling ensembles in TAC2017. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2019 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
26. Dandala B, Mahajan D, Devarakonda M. IBM Research system at TAC 2017: adverse drug reactions extraction from drug labels. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
27. Sun J, Gu X, Ding C, Li C, Li Y, Li S, et al. BUPT-PRIS system for TAC 2017 event nugget detection, event argument linking and ADR tracks. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017,. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
28. Tiftikci M, Özgür A, He Y, Hur J. BUPT-PRIS System for TAC 2017 Event Nugget Detection, Event Argument Linking and ADR Tracks. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
29. Xu J, Lee HJ, Ji Z, Wang J, Wei Q, Xu H. UTH CCB system for adverse drug reaction extraction from drug labels at TAC-ADR 2017. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
30. Pawar S, Palshikar G, Bhattacharyya P, Ramrakhiyani N, Gupta S, Varma V. TCS Research at TAC 2017: Joint extraction of entities and relations from drug labels using an ensemble of neural networks. In: Proceedings of the 2017 Text Analysis Conference, TAC 2017. 2017 Presented at: Proceedings of the 2017 Text Analysis Conference, TAC 2017; Nov 13-14, 2017; Gaithersburg, Maryland, USA.
31. Devlin J, Lee M, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv 2018:4805 [FREE Full text]
32. Du J, Xiang Y, Sankaranarayanapillai M, Zhang M, Wang J, Si Y, et al. Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning. J Am Med Inform Assoc 2021 Jul 14;28(7):1393-1400. [doi: 10.1093/jamia/ocab014] [Medline: 33647938]
33. Luo Y, Thompson WK, Herr TM, Zeng Z, Berendsen MA, Jonnalagadda SR, et al. Natural language processing for EHR-based pharmacovigilance: a structured review. Drug Saf 2017 Nov;40(11):1075-1089. [doi: 10.1007/s40264-017-0558-6] [Medline: 28643174]
34. Neves M, Leser U. A survey on annotation tools for the biomedical literature. Brief Bioinform 2014 Mar 18;15(2):327-340. [doi: 10.1093/bib/bbs084] [Medline: 23255168]
35. Tomanek K, Hahn U. Proceedings of the Linguistic Annotation Workshop; Suntec, Singapore2009. 2021 Presented at: Proceedings of the Linguistic Annotation Workshop; Suntec, Singapore2009; Proceedings of the Linguistic Annotation Workshop; Suntec, Singapore2009; Proceedings of the Linguistic Annotation Workshop; Suntec, Singapore2009 p. 112-115.
36. Deleger L, Li Q, Lingren T, Kaiser M, Molnar K, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc 2012;2012:144-153 [FREE Full text] [Medline: 23304283]
37. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296-298 [FREE Full text] [doi: 10.1197/jamia.M1733] [Medline: 15684123]
38. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Database issue):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
39. Spasic I, Krzeminski D, Corcoran P, Balinsky A. Cohort selection for clinical trials from longitudinal patient records: text mining approach. JMIR Med Inform 2019 Oct 31;7(4):e15980 [FREE Full text] [doi: 10.2196/15980] [Medline: 31674914]
40. 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]
41. Spasic I, Sarafraz F, Keane JA, Nenadic G. Medication information extraction with linguistic pattern matching and semantic rules. J Am Med Inform Assoc 2010;17(5):532-535 [FREE Full text] [doi: 10.1136/jamia.2010.003657] [Medline: 20819858]
42. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neur Inf Process Syst 2013 Dec;2:3111-3119 [FREE Full text]
43. Pennington J, Socher R, Manning C. GloVe: global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
44. Harris ZS. Distributional structure. Word 2015 Dec 04;10(2-3):146-162. [doi: 10.1080/00437956.1954.11659520]
45. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A. Attention is all you need. ArXiv.org. 2017. URL: https://arxiv.org/abs/1706.03762 [accessed 2021-12-12]
46. Wu Y, Schuster M, Chen Z, Le Q, Norouzi M, Macherey W. Google's neural machine translation system: bridging the gap between human and machine translation. ArXiv.org. 2016. URL: https://arxiv.org/abs/1609.08144 [accessed 2021-12-12]

47. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J. Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. 2016 Presented at: 12th USENIX Symposium on Operating Systems DesignImplementation; Nov 2 - 4, 2016; Savannah, GA. USA.
48. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
49. Grivas A, Alex B, Grover C, Tobin R, Whiteley W. Not a cute stroke: analysis of rule- and neural network-based information extraction systems for brain radiology reports. In: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis at the Conference on Empirical Methods in Natural Language Processing. 2020 Presented at: Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis at the Conference on Empirical Methods in Natural Language Processing; 2020; Louhi, Finland. [doi: 10.18653/v1/2020.louhi-1.4]
50. Wu S, Miller T, Masanz J, Coarr M, Halgrim S, Carrell D, et al. Negation's not solved: generalizability versus optimizability in clinical natural language processing. PLoS One 2014 Nov 13;9(11):e112774 [FREE Full text] [doi: 10.1371/journal.pone.0112774] [Medline: 25393544]
51. Rivera Zavala R, Martinez P. The impact of pretrained language models on negation and speculation detection in cross-lingual medical text: comparative study. JMIR Med Inform 2020 Dec 03;8(12):e18953 [FREE Full text] [doi: 10.2196/18953] [Medline: 33270027]
52. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller T. Does BERT need domain adaptation for clinical negation detection? J Am Med Inform Assoc 2020 Apr 01;27(4):584-591 [FREE Full text] [doi: 10.1093/jamia/ocaa001] [Medline: 32044989]
53. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. J Biomed Inform 2009 Oct;42(5):839-851 [FREE Full text] [doi: 10.1016/j.jbi.2009.05.002] [Medline: 19435614]
54. Sykes D, Grivas A, Grover C, Tobin R, Sudlow C, Whiteley W, et al. Comparison of rule-based and neural network models for negation detection in radiology reports. Nat Lang Eng 2020 Nov 18;27(2):203-224. [doi: 10.1017/s1351324920000509]
55. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform 2020 Mar 31;8(3):e17984 [FREE Full text] [doi: 10.2196/17984] [Medline: 32229465]
56. Kahle P, Colutto S, Hackl G, Mühlberger G. Transkribus - A service platform for transcription, recognition and retrieval of historical documents. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR); Kyoto, Japan. 2017 Presented at: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR); Kyoto, Japan; Nov 9-15, 2017; Kyoto, Japan. [doi: 10.1109/icdar.2017.307]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BioBERT: Bidirectional Encoder Representations from Transformers for biomedical text mining
BOW: bag of words
CRF: conditional random field
CTR: Center for Trials Research
CTU: clinical trial unit
CUI: concept unique identifier
FN: false negative
FP: false positive
LSTM: long short-term memory
NER: named entity recognition
NN: neural network
PR: precision-recall
SAE: serious adverse event
TP: true positive
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 09.03.21; peer-reviewed by B Alex, M Burns, S Shams; comments to author 07.06.21; revised version received 01.08.21; accepted 14.11.21; published 24.12.21.

Please cite as:
Chopard D, Treder MS, Corcoran P, Ahmed N, Johnson C, Busse M, Spasic I
Text Mining of Adverse Events in Clinical Trials: Deep Learning Approach
JMIR Med Inform 2021;9(12):e28632
URL: https://medinform.jmir.org/2021/12/e28632
doi:10.2196/28632
PMID:34951601

Original Paper

Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model

Hong Wu[1], MS; Jiatong Ji[2], MS; Haimei Tian[3], MS; Yao Chen[1], MS; Weihong Ge[4], MS; Haixia Zhang[4], PhD; Feng Yu[2], PhD; Jianjun Zou[5], PhD; Mitsuhiro Nakamura[6], PhD; Jun Liao[1], PhD

[1]School of Science, China Pharmaceutical University, Nanjing, China

[2]School of Basic Medicine and Clinical Pharmacy, China Pharmaceutical University, Nanjing, China

[3]School of Computer Engineering, Jinling Institute of Technology, Nanjing, China

[4]Department of Pharmacy, Nanjing Drum Tower Hospital, Nanjing, China

[5]Department of Clinical Pharmacology, Nanjing First Hospital, Nanjing Medical University, Nanjing, China

[6]Laboratory of Drug Informatics, Gifu Pharmaceutical University, Gifu, Japan

Corresponding Author:
Jun Liao, PhD
School of Science
China Pharmaceutical University
#639 Longmian Avenue
Jiangning District
Nanjing, 211198
China
Phone: 86 13952040425
Email: liaojun@cpu.edu.cn

Abstract

Background: With the increasing variety of drugs, the incidence of adverse drug events (ADEs) is increasing year by year. Massive numbers of ADEs are recorded in electronic medical records and adverse drug reaction (ADR) reports, which are important sources of potential ADR information. Meanwhile, it is essential to make latent ADR information automatically available for better postmarketing drug safety reevaluation and pharmacovigilance.

Objective: This study describes how to identify ADR-related information from Chinese ADE reports.

Methods: Our study established an efficient automated tool, named BBC-Radical. BBC-Radical is a model that consists of 3 components: Bidirectional Encoder Representations from Transformers (BERT), bidirectional long short-term memory (bi-LSTM), and conditional random field (CRF). The model identifies ADR-related information from Chinese ADR reports. Token features and radical features of Chinese characters were used to represent the common meaning of a group of words. BERT and Bi-LSTM-CRF were novel models that combined these features to conduct named entity recognition (NER) tasks in the free-text section of 24,890 ADR reports from the Jiangsu Province Adverse Drug Reaction Monitoring Center from 2010 to 2016. Moreover, the man-machine comparison experiment on the ADE records from Drum Tower Hospital was designed to compare the NER performance between the BBC-Radical model and a manual method.

Results: The NER model achieved relatively high performance, with a precision of 96.4%, recall of 96.0%, and F1 score of 96.2%. This indicates that the performance of the BBC-Radical model (precision 87.2%, recall 85.7%, and F1 score 86.4%) is much better than that of the manual method (precision 86.1%, recall 73.8%, and F1 score 79.5%) in the recognition task of each kind of entity.

Conclusions: The proposed model was competitive in extracting ADR-related information from ADE reports, and the results suggest that the application of our method to extract ADR-related information is of great significance in improving the quality of ADR reports and postmarketing drug safety evaluation.

(JMIR Med Inform 2021;9(12):e26407) doi:10.2196/26407

KEYWORDS

deep learning; BERT; adverse drug reaction; named entity recognition; electronic medical records

Introduction

Adverse drug reactions (ADRs) are a significant factor influencing the efficacy and safety of drugs and sometimes may even be life-threatening [1]. These safety problems are recorded as adverse drug events (ADEs) and reported to a special system like the Spontaneous Reporting System, which receives information from a wide range of sources, such as hospitals, small clinics, pharmacies, drug manufacturers, surveillance departments, and individuals [2]. Consequently, collecting and analyzing the ADEs that were recorded in the ADR reports have provided important content for drug safety supervision [3]. Conventional utilization of ADR reports mainly focuses on direct statistical analyses of structured sections [4,5], while the free-text section is quite underutilized because of the unstructured format. The unstructured part mainly describes the occurrence process of ADRs, which is a reference for supervisors to evaluate the potential ADRs. It involves a large amount of manual reading and a judgment process in the review step, which reduces the efficiency of evaluation and increases errors. Therefore, developing an automatic extraction tool to extract unstructured ADR-related information from Chinese ADE records is essential to improve the quality of ADR reports and postmarketing drug safety evaluation.

Named entity recognition (NER) is the main task of information extraction, in addition to natural language processing (NLP), the aim of which is to convert the unstructured contents into structured information. In the field of NLP, Word2Vec and other word vector methods [6-8] have been used for a long time to encode the text, which may bring only limited improvement to subsequent NLP tasks and fail to solve the polysemy problems [9,10]. Recently, numerous pretraining language models [11-13] have been proposed one after another, and Bidirectional Encoder Representations from Transformers (BERT) can greatly improve the performance of domain-related NLP tasks when it is fine-tuned with specific field datasets. BERT for Biomedical Text Mining (BioBERT) [14] was pretrained on large-scale biomedical corpora, which outperforms BERT on biomedical NER tasks, biomedical relation extraction tasks, and biomedical question answering tasks. And clinical NER (CNER) [15] also pretrained the BERT model on a large number of Chinese clinical literature sources crawled from the internet. Considering the background of ADRs in this study, we also collected a dataset of ADRs, and BERT was fine-tuned on this large unlabeled Chinese ADR-related corpus. As for the NER tasks, from the early dictionary-based [16] and rule-based method [17] to the traditional machine learning method [18] and then to the deep learning–based method, bidirectional long short-term memory (bi-LSTM) and conditional random field (CRF) have been widely used in the NER tasks. Wei et al [19] fused the results of CRF with those of bidirectional recurrent neural network (bi-RNN) by support vector machine and finally obtained a higher F1 score than those from CRF or bi-RNN models alone. The hybrid model of LSTM and CRF was proposed by Lample et al [20] in 2016, and its outstanding performance in many NER studies has made it the most popular NER model in recent years.

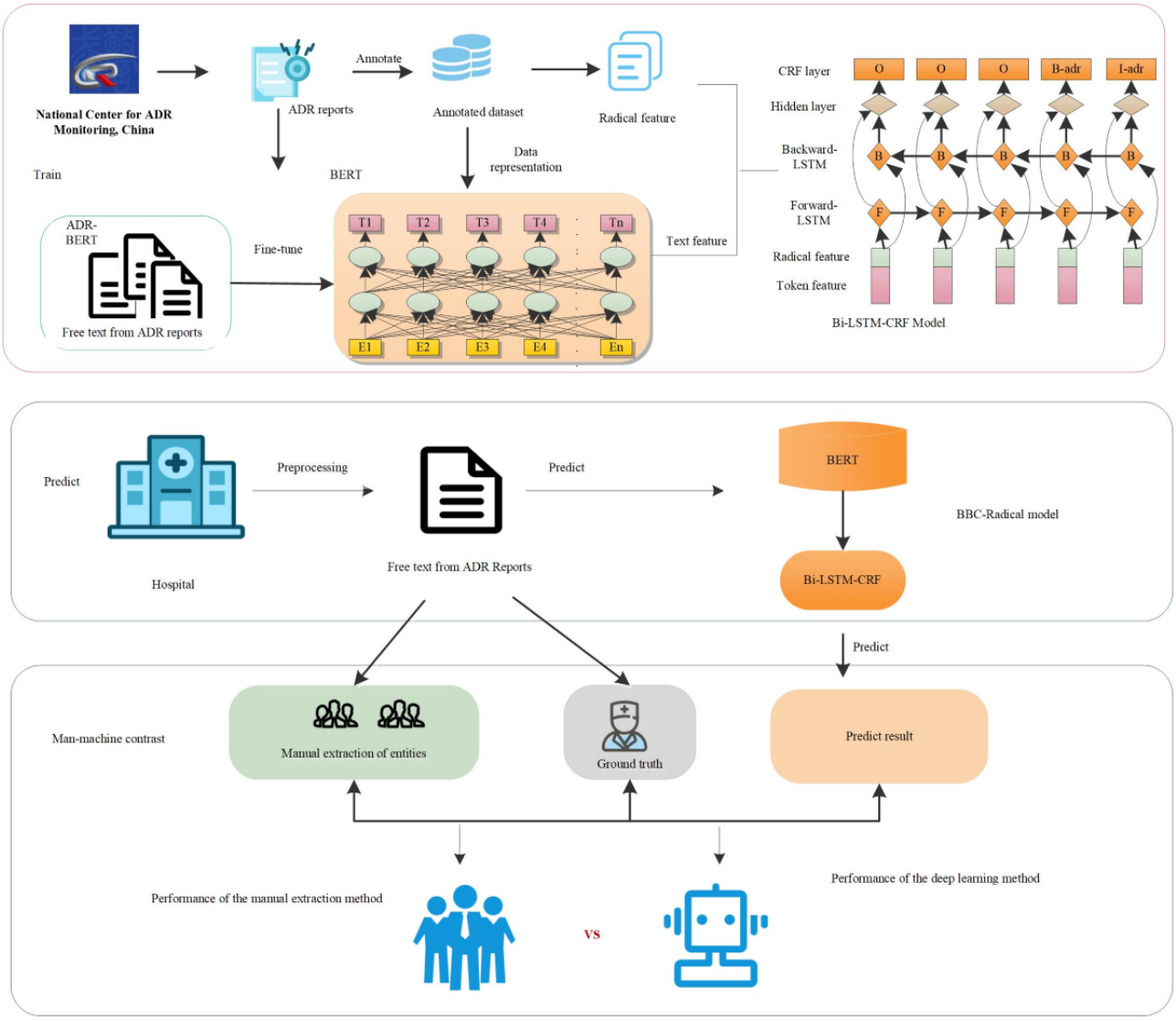
Consequently, in our study, we created a novel model, BERT-Bi-LSTM-CRF-Radical (BBC-Radical), that took token features and radical features as input and accurately recognized target entities in the sentence with the Bi-LSTM-CRF model. In order to better verify the performance of the model in the real world, we designed a Man-Machine comparison experiment based on the ADEs recorded by the Drum Tower Hospital from 2016 to 2019. We found that our method had excellent performance and efficiency (precision: 87.2%; recall: 85.7%; F1 score: 86.4%) versus manual method (precision: 86.1%; recall: 73.8%; F1 score: 79.5%). The automatically extracted ADR-related entities can further jointly serve as resources for ADR evaluation. On the whole, our study presented a novel method to identify ADR-related information from Chinese ADE reports.

Methods

Study Components

In our study, the model conducted the NER task in the free-text section of the Chinese ADE report from 2010-2016 from the Jiangsu province ADR Monitoring Center. According to the original content and the structure characteristics of the ADR cases and combined with other related research corpus annotation process, we established annotation rules and a tool for this study to recognize the entities and entity relationships between parts of the corpus annotation. In addition, the Man-Machine comparison experiment based on the ADE recorded by the Drum Tower Hospital was conducted to verify the extrapolation and robustness of the new data in the model. Figure 1 shows the pipeline of our study. The whole study can be divided into 3 parts: (1) training an NER model. The data representation model based on the BERT and the combination of token features (pink boxes) and radical features (green boxes) were fed into the Bi-LSTM-CRF model. Then, (2) the model performance was verified with real external data, and (3) a Man-Machine comparison experiment was designed to compare the efficiency and accuracy of NER tasks with a manual extraction method and a deep learning method.

Figure 1. The pipeline in our study; when training a named entity recognition (NER) mode, the data representation model based on the Bidirectional Encoder Representations from Transformers (BERT) model and the combination of token features (pink boxes) and radical features (green boxes) were fed into the bidirectional long short-term memory-conditional random field (bi-LSTM-CRF) model. ADR: adverse drug reaction, BBC-Radical: BERT-Bi-LSTM-CRF-Radical.



Dataset and Data Annotation

An ADR report can commonly be divided into 2 parts: structured section and free-text section. The data we used in this paper were from the unstructured section of Chinese ADR reports from the ADR monitoring center of Jiangsu Province in 2010-2016. The free-text section of a Chinese ADE report is the narrative content of the ADE procedure, commonly consisting of the process, solutions, and results of ADEs, along with the reasons for the medications being used to generate or degenerate the ADEs, in the form of one or more sentences, which may include some information that has not been recorded in the structured section. In this way, we applied NER technologies to extract entities automatically from these texts, which can be an auxiliary tool for ADR evaluation.

We manually annotated 24,890 cases from the free-text section from Chinese ADR reports, which have been described in [21]. To cover most of the cases, only 3 entities (“Reason,” “Drug,” and “ADR”) were annotated, with some other entities of low frequency not taken into consideration. The annotation rules and examples of entities are shown in Table 1.

Table 1. The definition and annotation rules and examples of entity annotation.

Entities and annotation rules	Examples
Reason[a]	
Symptoms or disease states associated with drug use	Diabetes, fever
Treatment involved with drug use	Postoperative fever
Drug[b]	
Generic names of medications	Levofloxacin
Trade names of medications	Lipitor
Abbreviations of medications	10% GS, 0.9% NS
ADR[a,c]	
Adverse reactions during or after medication	Bellyache

[a]Reference for the disease and adverse reaction definitions and classifications is the international Medical Dictionary for Regulatory Activities (MedDRA).
[b]"Drug" entity contains the generic name, trade name, abbreviation, and dosage form adjacent to the drug.
[c]ADR: adverse drug reaction.

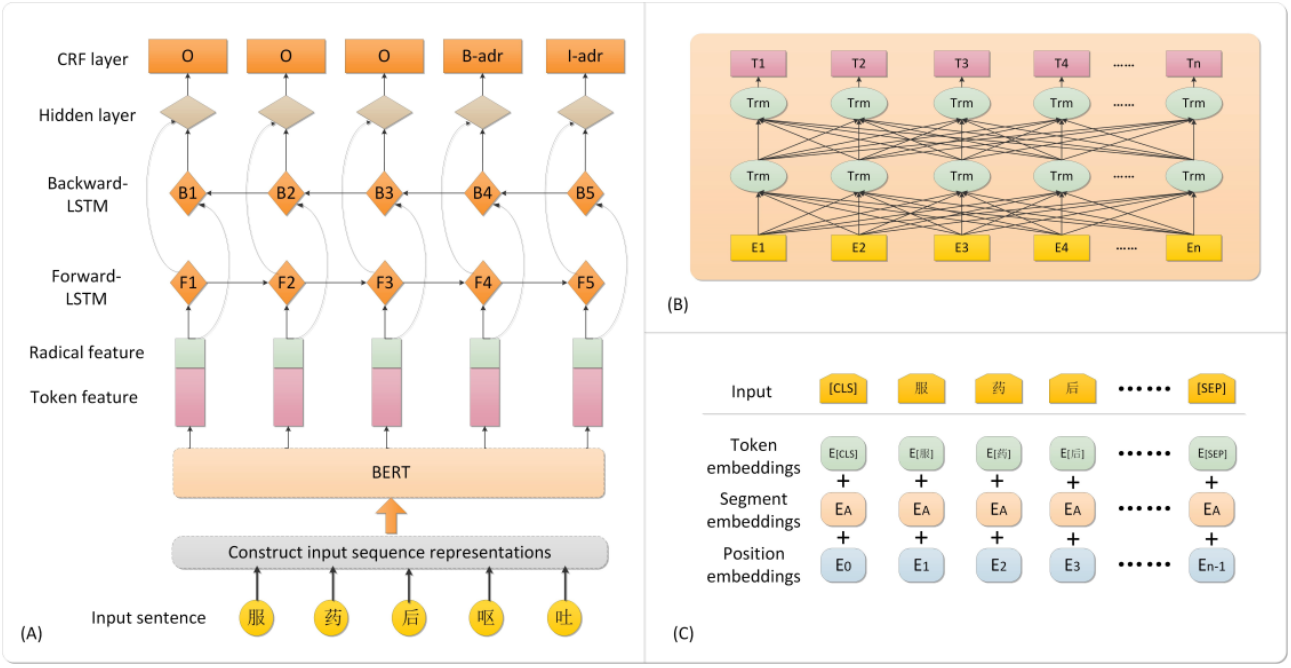
Input Representation of NER

To improve the efficiency of annotating work, the labeled cases were annotated by an efficient tool [21,22]. We used a special [CLS] at the beginning of the sentence, used [SEP] to separate segments or denote the end of the sequence, and added [PAD] tokens at the end of the sentences to make their lengths equal to the maximum sequence length. Finally, 24,890 valid annotated cases were obtained, including 147,451 entities in the Chinese ADR reports.

BBC-Radical Method

Figure 2B shows the data representation model based on BERT that takes the corresponding token, segment, and position embeddings of each word as inputs in Figure 2C. The contextual embeddings for each token can be obtained from the output of the BERT model, which is the token feature input of the next NLP task.

Figure 2. (A) Architecture diagram of our proposed model, in which the combination of token features (pink boxes) and radical features (green boxes) were fed into the bidirectional long short-term memory-conditional random field (Bi-LSTM-CRF) model; (B) data representation model based on the Bidirectional Encoder Representations from Transformers (BERT) model, in which the sequence of [E_1, E_2, E_3 … E_n] in the yellow boxes is the input to the BERT model and the green ellipses represent the Transform blocks; and (C) construction of input sequence representations for the BERT, in which the input is composed of token embedding, segment embedding, and position embedding.



Owing to the fact that the corpus we used in this study is highly domain-specific, we also collected nearly 461,930 ADE records from the ADR Monitoring Center of Jiangsu Province ranging from 2010 to 2016 to improve the precision of domain-specific word representation. These records were mainly from medical personnel in hospitals and pharmacies and from follow-up

records from pharmaceutical companies. The diversity of submitting agencies and reporters enriched the sample and made the language characteristics of the data sources more complex. To fine-tune BERT, we first generated a pretraining data (a tfrecord file) file with the clinical text. Then, we pretrained our fine-tuned BERT model on the pretraining file from the existing BERT checkpoint of the original language model ($BERT_{BASE\text{-}Chinese\text{-}uncased}$). Once the fine-tuning process was completed, we got a TensorFlow model that was transformed to a PyTorch model for further NER tasks.

Radical is a common form extracted from many Chinese characters, so that these characters not only have the basis of classification in the form but also become a common genus in the meaning of the word, which is helpful for people to summarize the meaning of the word. Therefore, the meaning of radicals is very important for people to grasp the meaning of a word. Moreover, the radical features of Chinese characters have been widely used to enhance different Chinese NLP tasks in recent years [23-25]; consequently, in addition to considering Chinese characters themselves, we also considered applying radical features to the model. The overall network architecture of our NER model is shown in the Bi-LSTM-CRF in Figure 2A. In our study, each token in our sequence was fed into the fine-tuned BERT model to train for the data representation of the whole sequence. After obtaining a representation of the entire sequence [T_1, T_2, T_3, …T_n], we looked for the radical of each word in the sequence and initialized each radical with random values to indicate the radical feature. The concatenation $x = [w\_1, w\_2, w\_3, \ldots w\_4]$ of the word vectors and the radical vectors were fed into the Bi-LSTM model, and the context vectors learned by forward and backward LSTM layers were then transmitted into the CRF layer to compute the corresponding probability values and to simultaneously predict tags. The details of our NER method in the Bi-LSTM-CRF are shown in Multimedia Appendix 1. We also implemented 3 baselines for the NER task, as follows:

1. CRF + + is a well-known open-source tool for CRF that is also the CRF tool with the best comprehensive performance at present.
2. The Bi-LSTM-CRF model that takes the input representation trained by Word2Vec as input was used as a baseline.
3. The combined model, BBC-Radical model without fine-tuning BERT on domain-specific corpus, was also used as a baseline (BERT + Bi-LSTM-CRF-Radical). The model architectures and experimental settings were the same as in our proposed model.

Results

Experimental Settings

All the models were trained on an NVIDIA Tesla V100 GPU with 768 GB of memory using the PyTorch framework. The longest length of a sentence can be set to 512 in the fine-tuned BERT of our NER model. To maintain the complete information from the sentences, the excess part was split into another sentence once the length exceeded 512 tokens, until all the segmented sentences could satisfy the length constraint. We trained the model with a batch size of 16, the hidden unit of bi-LSTM was 128, and we also used radical embedding, which was initialized with 20 random values. We also set the initial learning rate as 510^{-5} in the Adam optimizer.

Evaluation Metrics

The results were measured using a micro-averaged F1 score = $(2PR)/(P + R)$, where P denotes precision and R represents recall. In our research, we followed the strict matching rule that was defined at the start and end boundaries, and the extraction result referred to the same entity types as the ground truth.

Findings

In the NER task, 15,000 and 8000 cases were randomly selected separately from the annotated cases as the training set and testing set, respectively, and the remaining 1890 cases were considered the validation set, which was used to verify the generalization ability of the model in the training process. In order to better evaluate the model, we ran our proposed model and the third baseline model 10 times, keeping all the other parameters the same except the sampled training data. The average value of each valuation metric was used to show the prediction results in Table 2.

Table 2. Overall concept extraction performances from the free-text section of Chinese adverse drug reaction (ADR) reports.

Model	Precision (%), mean (SD)	Recall (%), mean (SD)	F1 score (%), mean (SD)
CRF[a] ++ [21]	94.4 (0.32)	93.1 (0.28)	93.9 (0.08)
Word2Vec + Bi-LSTM[b]-CRF [21]	94.6 (0.33)	94.1 (0.30)	94.4 (0.29)
BERT[c] + Bi-LSTM-CRF-Radical	95.2 (0.07)	95.2 (0.07)	95.2 (0.06)
Fine-tuning BERT + Bi-LSTM-CRF	96.0 (0.05)	95.5 (0.08)	96.0 (0.06)
Fine-tuning BBC[d]-Radical	96.4 (0.04)	96.0 (0.03)	96.2 (0.04)

[a]CRF: conditional random field.

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

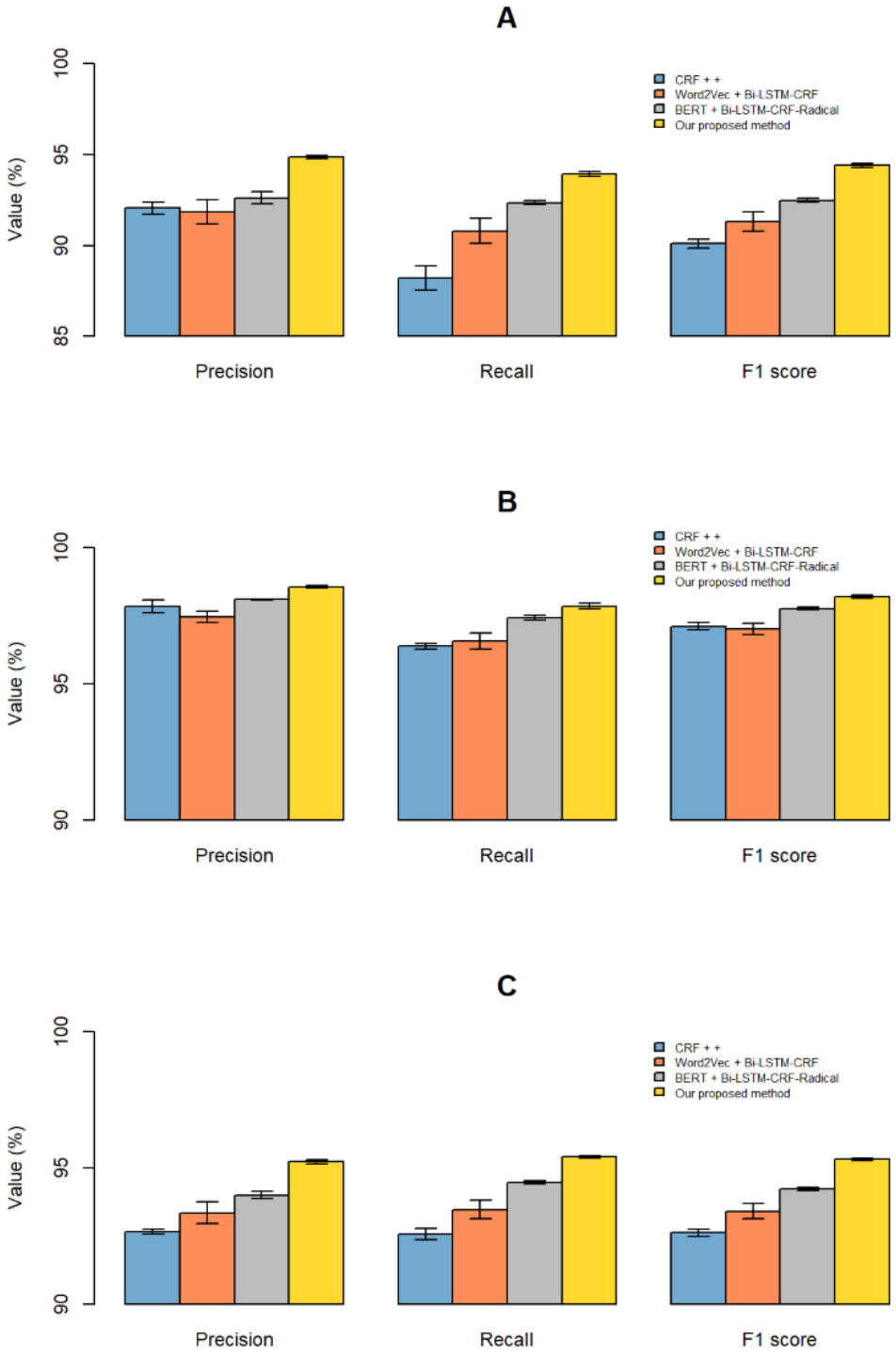
[c]BERT: Bidirectional Encoder Representations from Transformers.

[d]BBC: BERT-Bi-LSTM-CRF.

CRF is a probabilistic structure model by marking and segmenting sequence data. The obvious disadvantage of the word vector model represented by Word2vec is that it is context-free, which results in the same word having the same meaning in different contexts. Consequently, neither the CRF++ nor the second baseline model did very well in our NER task. The third row in Table 2 represents the model combining the original BERT and Bi-LSTM-CRF-Radical models, and the BERT model in our proposed BBC-Radical model was fine-tuned on the domain-specific corpus. The results in the fourth row of Table 2 also show the contribution of radical embedding in the NER task. The proposed model in our research achieved an F1 score of 96.2%, which outperforms the 4 baseline models. The results showed that BERT plays an important role in capturing more text information, and our pre-trained BERT on a specific domain can significantly improve the performance of entity extraction.

Our proposed method outperformed the other methods for all entity types, and the entity of “Drug” achieved the highest F1 score, while the entity of “Reason” achieved the lowest (Figure 3). This can be implied from the definitions of each entity, in which the entity of “Reason” included not only conventional diseases and symptoms but also some other treatments involved with drug use, along with their body parts and adjoining adjectives, while the definitions of “Drug” and “ADR” were relatively simpler. Because the definitions or annotations of the rules were more varied, the error rate of the model was relatively high. The overlap of the concepts of different kinds of entities is another reason for the false recognition between “Reason” and “ADR.” For instance, the entity of “Reason” was always recorded in ADR reports with the colloquial expressions of symptoms. And it was hard to recognize the “Reason” of “anorexia” when it was recorded as “never feel like eating.” As for the entity of “Drug,” we found that it was difficult to identify infrequently used trade names, some English abbreviations, and some traditional Chinese medicines that are composed of peculiar characters.

Figure 3. Precision, recall, and F1 score for each kind of entity: (A) reason, (B) drug, and (C) adverse drug reaction (ADR). CRF: conditional random field; BERT: Bidirectional Encoder Representations from Transformers; bi-LSTM: bidirectional long short-term memory.



Validation Results

The Man-Machine Contrast on the External Validation Dataset

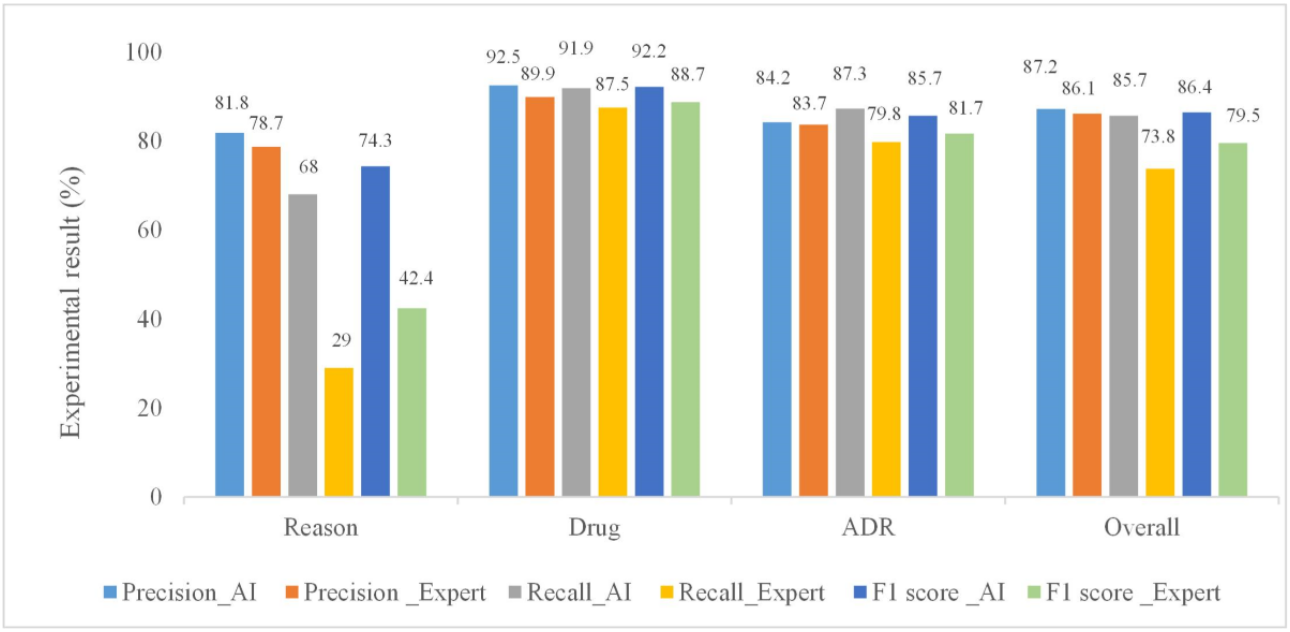
As the frontier direction of artificial intelligence research, man-machine confrontation technology has always been a hot spot of artificial intelligence research. Research of artificial intelligence, mainly in the form of man-machine confrontation, provides an excellent experimental environment and verification method for exploring the internal growth mechanism and key technical principles of machine intelligence. The whole process can not only make the machine serve humans more intelligently but also free humans from some complex tasks. For further validation, we selected 2479 ADE reports recorded by physicians from The Drum Tower Hospital, School of Medicine, Nanjing University from 2016 to 2019 and validated our proposed model to conduct NER experiments on the descriptive texts of adverse events. After professional training, the ground truth for the 2479 cases was produced by 10 students majoring in hospital pharmacy, who spent 2 weeks for 4 rounds of annotating (including 1 round of pre-annotating, 2 rounds of

formal annotating, and 1 round of annotation correction). The rules for annotation were the same as those for our annotation training set. In order to further illustrate the advantages of the entity recognition model in this paper, we designed a Man-Machine comparison experiment. In reality, the hospital reports the ADEs to the ADR center every year, and then, the staff of the adverse reaction center needs to review and return reports that do not conform to the standard. Therefore, there are some differences in the recognition performance between the ADEs reported by the hospital and the data from the ADR monitoring center. As for the manual method, we invited 5 additional pharmacy students to participate in the experiment. Under the guidance of the ADR supervisor, the 5 students were required to complete the entity extraction of the assigned data by manual search within 2 weeks after training. Since manual entity extraction is time-consuming and laborious, we only had 2 rounds of marking, and finally, we obtained the results of manual entity extraction. The results of the Man-Machine comparison to the external validation data are shown in Multimedia Appendix 2.

Comparison of the Man-Machine Contrast Results

After the preprocessing step, the validation data were fed into our model for prediction, and the performance of the prediction results is provided in Figure 4 (light blue, gray, and blue bars). From the perspective of the entity category, when the target entity "Reason" was only defined as "Disease," the identification accuracy of the entity was relatively good [26], while our definition of "Reason" also contained other drug use–related treatments and symptoms. Tao et al [27] performed the NER task of "Reason" and other medicine-related entities, and the resulting F1 score for "Reason" was only 40.9%; our F1 score for "Reason" was 74.3%. Figure 4 (orange, yellow, and green bars) also shows the results of the manual extraction method. The manual method achieved a precision of 86.1%, recall of 73.8%, and F1 score of 79.5% for the NER task. Therefore, manual extraction of entities was not only inefficient, but also had low accuracy, especially the identification of the entity "Reason." The accuracy of "Drug" entity extraction was relatively high due to the normative name of the "Drug" entity in each experiment, while the extraction of the "Reason" and "ADR" entities was also negatively affected by nonstandard documentation.

Figure 4. Comparison of the Man-Machine contrast. ADR: adverse drug reaction.



In Figure 4, the F1 score is only 42.4%, and recall is extremely low in the recognition of "Reason," far lower than with our deep learning method. A possible reason could be that many entities were not recognized when the entity of "Reason" was identified manually because of limited human attention and accuracy. At the same time, by comparison, we found that the F1 score of the BBC-Radical model for the recognition of other entities was also much better than the manual recognition method. As a special machine learning method, deep learning can automatically extract features from data samples, which reduces the process of constructing artificial features and has more advantages for processing large data sets.

Discussion

Principal Findings

In our study, we developed a domain-specific NER method on Chinese ADE records. The extraction of biomedical entities and their relationships from texts is of great application value to biomedical research. Accurately extracting entity information from free text in Chinese ADE reports with NER methods in daily practical work can greatly simplify the approval work of staff in the ADR monitoring center and improve the quality of ADE reports. In addition to using medical reports for detecting ADRs, it has been proposed to use data from social media [28], since users tend to discuss their illnesses, treatments, and prescribed medications and their effects on social media

platforms. For example, when Cocos et al [29] and Xie et al [30] extracted ADR entities from social media using dictionary matching, using a CRF and bi-LSTM model, respectively, they helped reduce the limitations of passive reporting systems. Also, the automatic detection of chemical references in biomedical literature is an essential step for further biomedical text mining and has recently received considerable attention. In addition to using a single model for training, Zeng et al [31] and Luo et al [32] achieved high F1 scores when they integrated bi-LSTM and CRF to extract the drug entity and chemical substance entity, respectively, from the text. The performance of the baseline model using a single CRF + + also proved that the single CRF model was inferior to the BI-LSTM-CRF model. Due to the excellent performance of the hybrid model with bi-LSTM and CRF, the hybrid model architecture with bi-LSTM and CRF was also applied in the entity extraction layer of our proposed model.

Most NLP tasks based on deep learning can be divided into the following 3 modules: data processing, text representation, and a task-specific model. Word2Vec, GloVe, and BERT are excellent models of text representation that are widely used in different NER models. Chen et al [21] obtained a high F1 score when Word2Vec and Bi-LSTM-CRF were used to extract the named entities from the free-text section of Chinese ADR Reports. However, Chen et al [21] applied Word2Vec in the input layer to generate the data representation, which would bring only limited improvement to subsequent NLP tasks and failed to solve the polysemy problem. In Chinese text, Zhang et al [33] extracted breast cancer–related entities with a pretrained BERT model that was trained on a large-scale, unlabeled corpus of Chinese clinical text. However, their pretrained BERT on this domain was aimed at breast cancer, not general medical records. In our study, we also established a deep neural network algorithm based on a domain-specific BERT model, and our model proved the competitive performance of NER on ADE text in the setting of the same training set with a higher F1 score. From the results of the Man-Machine comparison experiment, our proposed method achieved a high degree of agreement with ground truth. Moreover, the method proposed in this paper was superior to manual extraction in the accuracy and speed of NER.

Furthermore, the use of NER in NLP technology can achieve the target entity in automatic extraction from free text, and the extraction of information can be further used in statistical analysis, such as knowledge base–building tasks. Besides that, the model can be used to automatically extract ADR-related information from electronic medical records or other relevant texts to further supplement the information contained in ADR reports.

Conclusion

In this study, we explored an NER task on Chinese ADR reports, with an optimized deep learning method of BBC-Radical, which took radical features of each token, token features obtained from the fine-tuned BERT model as the input, and Bi-LSTM-CRF as the feature extract model. The performance of our model was compared with other baseline models on the same dataset, and the experimental results indicated that the BBC-Radical model outperformed other models and obtained a competitive F1 score of 96.2%. Moreover, in the Man-Machine comparison experiment, our method had an absolute advantage over the manual extraction method in terms of time, efficiency, and accuracy. This study conducted a domain-specific NER task in Chinese ADE records, which may play a role in promoting ADR evaluation and postmarketing evaluation of drug safety.

Acknowledgments

This work was supported by the 2017-2018 annual scientific research project of the Jiangsu Food and Drug Administration (NO 20170308), “Double First-Class” University project (NO CPU2018GY19), and National Natural Science Foundation of China (NO 81673511). We also acknowledge the computing support from the High Performance Computing Center of the China Pharmaceutical University.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Details of our named entity recognition method.
[DOCX File , 20 KB - medinform_v9i12e26407_app1.docx]

Multimedia Appendix 2
Results from the external validation data.
[DOCX File , 17 KB - medinform_v9i12e26407_app2.docx]

References

1. Li X, Lin X, Ren H, Guo J. Ontological organization and bioinformatic analysis of adverse drug reactions from package inserts: Development and usability study. J Med Internet Res 2020 Jul 20;22(7):e20443 [FREE Full text] [doi: 10.2196/20443] [Medline: 32706718]

2. Pal SN, Duncombe C, Falzon D, Olsson S. WHO strategy for collecting safety data in public health programmes: complementing spontaneous reporting systems. Drug Saf 2013 Feb 18;36(2):75-81 [FREE Full text] [doi: 10.1007/s40264-012-0014-6] [Medline: 23329541]
3. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. The Lancet 2000 Oct;356(9237):1255-1259. [doi: 10.1016/s0140-6736(00)02799-9]
4. Pageot C, Bezin J, Smith A, Arnaud M, Salvo F, Haramburu F, French Network of Pharmacovigilance Centres. Impact of medicine withdrawal on reporting of adverse events involving therapeutic alternatives: A study from the French Spontaneous Reporting Database. Drug Saf 2017 Nov 29;40(11):1099-1107. [doi: 10.1007/s40264-017-0561-y] [Medline: 28664354]
5. Schwan S, Sundström A, Stjernberg E, Hallberg E, Hallberg P. A signal for an abuse liability for pregabalin--results from the Swedish spontaneous adverse drug reaction reporting system. Eur J Clin Pharmacol 2010 Sep 19;66(9):947-953. [doi: 10.1007/s00228-010-0853-y] [Medline: 20563568]
6. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. Cornell University. 2013. URL: https://arxiv.org/abs/1310.4546 [accessed 2021-11-08]
7. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. Stanford. 2014. URL: https://nlp.stanford.edu/projects/glove/ [accessed 2021-11-08]
8. 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]
9. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. 2016 Presented at: 12th USENIX conference on Operating Systems Design and Implementation; November 2-4, 2016; Savannah, GA.
10. Deng L. Deep Learning: Methods and Applications. FNT in Signal Processing 2013;7(3-4):197-387. [doi: 10.1561/2000000039]
11. Cui Y, Che W, Liu T, Qin B, Yang Z. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. Audio Speech Lang. Process 2019:1-1. [doi: 10.1109/taslp.2021.3124365]
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Cornell University. 2018. URL: https://arxiv.org/abs/1810.04805 [accessed 2021-11-08]
13. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. URL: https://tinyurl.com/49576n96 [accessed 2021-11-08]
14. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
15. Li X, Zhang H, Zhou X. Chinese clinical named entity recognition with variant neural structures based on BERT methods. J Biomed Inform 2020 Jul;107:103422 [FREE Full text] [doi: 10.1016/j.jbi.2020.103422] [Medline: 32353595]
16. Ekbal A, Saha S. Simultaneous feature and parameter selection using multiobjective optimization: application to named entity recognition. Int. J. Mach. Learn. & Cyber 2014 Jul 6;7(4):597-611. [doi: 10.1007/s13042-014-0268-7]
17. Oudah M, Shaalan K. NERA 2.0: Improving coverage and performance of rule-based named entity recognition for Arabic. Nat. Lang. Eng 2016 May 06;23(3):441-472. [doi: 10.1017/s1351324916000097]
18. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform 2009 Oct;42(5):905-911 [FREE Full text] [doi: 10.1016/j.jbi.2008.12.012] [Medline: 19535010]
19. Wei Q, Chen T, Xu R, He Y, Gui L. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. Database (Oxford) 2016 Oct 24;2016:baw140 [FREE Full text] [doi: 10.1093/database/baw140] [Medline: 27777244]
20. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. Cornell University. 2016. URL: https://arxiv.org/abs/1603.01360 [accessed 2021-11-08]
21. Chen Y, Zhou C, Li T, Wu H, Zhao X, Ye K, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. J Biomed Inform 2019 Aug;96:103252 [FREE Full text] [doi: 10.1016/j.jbi.2019.103252] [Medline: 31323311]
22. cpuchenyao / NER_RE_Annotation. GitHub. 2018 Nov 26. URL: https://github.com/cpuchenyao/NER_RE_Annotation [accessed 2021-11-08]
23. Peng H, Cambria E, Zou X. Radical-Based Hierarchical Embeddings for Chinese Sentiment Analysis at Sentence Level. Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference. 2017. URL: https://sentic.net/radical-embeddings-for-chinese-sentiment-analysis.pdf [accessed 2021-11-08]
24. Shao Y, Hardmeier C, Tiedemann J, Nivre J. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF. Cornell University. 2017. URL: https://arxiv.org/abs/1704.01314 [accessed 2021-11-08]
25. Shi X, Zhai J, Yang X, Xie Z, Liu C. Radical Embedding: Delving Deeper to Chinese Radicals. 2015. URL: https://aclanthology.org/P15-2098.pdf [accessed 2021-11-08]
26. Pons E, Becker BF, Akhondi SA, Afzal Z, van Mulligen EM, Kors JA. Extraction of chemical-induced diseases using prior knowledge and textual information. Database (Oxford) 2016 Apr 14;2016:baw046 [FREE Full text] [doi: 10.1093/database/baw046] [Medline: 27081155]

27. Tao C, Filannino M, Uzuner O. Prescription extraction using CRFs and word embeddings. J Biomed Inform 2017 Aug;72:60-66 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.002] [Medline: 28684255]
28. Arnoux-Guenegou A, Girardeau Y, Chen X, Deldossi M, Aboukhamis R, Faviez C, et al. The adverse drug reactions from patient reports in social media project: Protocol for an evaluation against a gold standard. JMIR Res Protoc 2019 May 07;8(5):e11448 [FREE Full text] [doi: 10.2196/11448] [Medline: 31066711]
29. Cocos A, Fiks A, Masino A. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. J Am Med Inform Assoc 2017 Jul 01;24(4):813-821 [FREE Full text] [doi: 10.1093/jamia/ocw180] [Medline: 28339747]
30. Xie J, Liu X, Dajun Zeng D. Mining e-cigarette adverse events in social media using Bi-LSTM recurrent neural network with word embedding representation. J Am Med Inform Assoc 2018 Jan 01;25(1):72-80 [FREE Full text] [doi: 10.1093/jamia/ocx045] [Medline: 28505280]
31. Zeng D, Sun C, Lin L, Liu B. LSTM-CRF for drug-named entity recognition. Entropy 2017 Jun 17;19(6):283. [doi: 10.3390/e19060283]
32. Luo L, Yang Z, Yang P, Zhang Y, Wang L, Lin H, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics 2018 Apr 15;34(8):1381-1388. [doi: 10.1093/bioinformatics/btx761] [Medline: 29186323]
33. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. Int J Med Inform 2019 Dec;132:103985. [doi: 10.1016/j.ijmedinf.2019.103985] [Medline: 31627032]

Abbreviations

ADE: adverse drug event
ADR: adverse drug reaction
BBC-Radical: BERT-Bi-LSTM-CRF-Radical
BERT: Bidirectional Encoder Representations from Transformers
bi-LSTM: bidirectional long short-term memory
bi-RNN: bidirectional recurrent neural network
BioBERT: BERT for Biomedical Text Mining
CNER: clinical NER
CRF: conditional random field
NER: named entity recognition
NLP: natural language processing

Edited by R Kukafka, G Eysenbach; submitted 10.12.20; peer-reviewed by KNB Nor Aripin, J Zheng; comments to author 07.03.21; revised version received 22.04.21; accepted 05.10.21; published 01.12.21.

Please cite as:
Wu H, Ji J, Tian H, Chen Y, Ge W, Zhang H, Yu F, Zou J, Nakamura M, Liao J
Chinese-Named Entity Recognition From Adverse Drug Event Records: Radical Embedding-Combined Dynamic Embedding–Based BERT in a Bidirectional Long Short-term Conditional Random Field (Bi-LSTM-CRF) Model
JMIR Med Inform 2021;9(12):e26407
URL: https://medinform.jmir.org/2021/12/e26407
doi:10.2196/26407
PMID:34855616

Original Paper

Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study

Nicolas Paris[1*], MSc; Antoine Lamer[1,2], PhD; Adrien Parrot[1*], MSc, MD

[1]InterHop, Paris, France

[2]Univ. Lille, CHU Lille, ULR 2694 - METRICS: Évaluation des Technologies de santé et des Pratiques médicales, Lille, France

[*]these authors contributed equally

Corresponding Author:
Nicolas Paris, MSc
InterHop
30 avenue du Maine
Paris, 75015
France
Phone: 33 3 20 62 69 69
Email: nicolas.paris@riseup.net

Abstract

Background: In the era of big data, the intensive care unit (ICU) is likely to benefit from real-time computer analysis and modeling based on close patient monitoring and electronic health record data. The Medical Information Mart for Intensive Care (MIMIC) is the first open access database in the ICU domain. Many studies have shown that common data models (CDMs) improve database searching by allowing code, tools, and experience to be shared. The Observational Medical Outcomes Partnership (OMOP) CDM is spreading all over the world.

Objective: The objective was to transform MIMIC into an OMOP database and to evaluate the benefits of this transformation for analysts.

Methods: We transformed MIMIC (version 1.4.21) into OMOP format (version 5.3.3.1) through semantic and structural mapping. The structural mapping aimed at moving the MIMIC data into the right place in OMOP, with some data transformations. The mapping was divided into 3 phases: conception, implementation, and evaluation. The conceptual mapping aimed at aligning the MIMIC local terminologies to OMOP's standard ones. It consisted of 3 phases: integration, alignment, and evaluation. A documented, tested, versioned, exemplified, and open repository was set up to support the transformation and improvement of the MIMIC community's source code. The resulting data set was evaluated over a 48-hour datathon.

Results: With an investment of 2 people for 500 hours, 64% of the data items of the 26 MIMIC tables were standardized into the OMOP CDM and 78% of the source concepts mapped to reference terminologies. The model proved its ability to support community contributions and was well received during the datathon, with 160 participants and 15,000 requests executed with a maximum duration of 1 minute.

Conclusions: The resulting MIMIC-OMOP data set is the first MIMIC-OMOP data set available free of charge with real disidentified data ready for replicable intensive care research. This approach can be generalized to any medical field.

(JMIR Med Inform 2021;9(12):e30970) doi:10.2196/30970

KEYWORDS

data reuse; open data; OMOP; common data model; critical care; machine learning; big data; health informatics; health data; health database; electronic health records; open access database; digital health; intensive care; health care

Introduction

Intensive care units (ICUs) are designed to provide comprehensive support to the most severely ill patients in a hospital [1]. Mortality is typically high among these patients, both during and after the hospital stay [2]. Understanding the effects of interventions on patient outcomes remains a challenge due to the heterogeneity of patients, complexity of disease, and variation in care patterns. Intensivists use a limited level of evidence to guide decision making [3], whereas ICUs are a high-density environment for data production.

With the increasing adoption of electronic health record (EHR) systems around the world leading to large amounts of clinical data [4] and the development of data mining, innovation through data reuse is likely to play an important role in clinical medicine [5]. Indeed, based on important medical information, expectations are to improve clinical outcomes and practices, enable personalized medicine and guide early warning systems, and also easily enroll a large, multicenter cohort, while minimizing costs [6,7].

The Medical Information Mart for Intensive Care (MIMIC)-III is a high-granularity data set of over 60,000 intensive care stays and 46,000 unique patients from 2 successive ICU systems at the Beth Israel Deaconess Medical Center in Boston, admitted from 2001 to 2012 [8]. It is the first ICU database available for free, and it has been intensively used in research, resulting in more than 300 international publications. However, its monocentric nature makes it difficult to generalize findings to other ICUs.

For Kahn et al [9], "Database modelling is the process of determining how data are to be stored in a database." It specifies data types, constraints, relationships, and metadata definitions and provides a standardized way to represent resources/data and their relationships. Some studies have shown that using a common data model (CDM) by standardizing the structure (data model) and concepts (terminological model) of the database allows larger-scale multicenter research and exploitation of rare diseases or rare events and catalyzes research by sharing practices, source code, and tools [10,11]. However, some studies have shown that the results are not fully reproducible from one CDM to another [12] or from one center to another [13]. Some approaches argue that keeping the local conceptual model [14] and the local structural model [15] leads to better results. On the one hand, keeping MIMIC in its specific form will not solve the limitation for multicenter research, but on the other hand, a fully standardized form would introduce other disadvantages, such as loss of data and lower computational performances. The ideal solution is probably in between to allow local or standardized analysis, depending on the research question.

The Observational Medical Outcomes Partnership (OMOP) CDM is a data model originally designed for multicenter research related to adverse drug events, which has been now extended to medical, laboratory, and genomic cases. OMOP provides structural and conceptual models relying on reference terminologies, such as Systematized Nomenclature of Medicine (SNOMED) for diagnostics, RxNORM for drugs, and Logical Observation Identifiers Names and Codes (LOINC) for laboratory results. Several examples of databases transformed into OMOP have been published [16-18], and OMOP stores more than half a billion patient records from around the world [19,20]. The OMOP conceptual model is based on a closure table pattern [21] capable of ingesting any simple, hierarchical, and also graph terminologies, such as SNOMED. In addition to local terminologies, OMOP defines and maintains a set of standard terminologies to be mapped unidirectionally (local to standard) by implementers. Although OMOP has proven its reliability [22], the concept mapping process is known to have an impact on results [23] and the application of the same protocol on different data sources leads to different results [13]. This shows the importance of keeping local terminologies so that local analysis is still possible. Previous preliminary work has been done on the translation of MIMIC into OMOP [24]. This work remains to be refined and updated for proper evaluation.

When comparing different CDMs [10,25], OMOP obtained the best results for completeness; integrity; flexibility; simplicity of integration; implementability for a wider coverage of the structural and conceptual model; a more systematic analysis, thanks to an analytical library and to visualization tools; and easier access to data through SQL queries. In terms of a conceptual approach, OMOP offers a broader set of standard concepts. In terms of a structural CDM, it is rigorous in how data should be loaded into specific tables, while other CDMs, such i2b2, are flexible with a general table that solves all data domains. This rigorous approach is necessary for standardization. Previous work has been performed to load MIMIC-III into i2b2 [26]; however, the work could not be finalized due to the tricky concept mapping to standard terminologies tasks. OMOP has the advantage of not making the terminology-mapping step mandatory by keeping the local codes accessible to analysts. Compared to the Fast Healthcare Interoperability Resources (FHIR) [27], OMOP performs better as a conceptual CDM because the FHIR resources currently do not specify the terminology to be used for most of the attributes. The OMOP relational model can be materialized in csv format and stored in any relational database, while the FHIR uses json files and needs some processing and higher skills to exploit. Among the above models, OMOP is the best candidate to overcome the MIMIC limitations mentioned earlier.

Our paper was guided by the 2 following objectives: (1) transforming MIMIC into OMOP in terms of the time needed, skills required, and quality of the result and (2) evaluating the resulting data set to support efficient, shareable, and real-time analysis.

Methods

Data

The majority of source code was implemented in PostgreSQL 9.6.9 (Postgres) because it is the primary support for the MIMIC database. It also allows the community to reproduce our work on limited resources without licensing costs and benefit from recent Postgres improvements in the data processing area. Some elaborated data transformations have been implemented as Postgres functions.

OMOP CDM version 5.3.3.1 (OMOP) tables were created from the provided scripts, with some changes documented in our scripts. OMOP defines 15 standardized CLINICAL data tables, 3 HEALTH system data tables, 2 HEALTH ECONOMICS data tables, 5 tables for DERIVED elements, and 12 tables for standardized VOCABULARY. The VOCABULARY tables were loaded from concepts downloaded from Athena [28], and the clinical and derived tables were loaded from MIMIC.

MIMIC-III version 1.4.21 (MIMIC) was also loaded into Postgres with the provided scripts. A subset of 100 patients over the 46,000 total MIMIC patients was selected based on their

broad representativeness in the database and was cloned into a second instance to serve as a light and representative development set.

Structural Mapping

The *structural mapping* aimed at moving the MIMIC data to the right place in OMOP, with some data transformations. It was organized into 3 phases: conception, implementation, and evaluation.

The *conception* phase consisted of looping over each MIMIC table and choosing an equivalent location in OMOP for each column. In general, both projects were appropriately documented, but in several cases, we needed some clarification from MIMIC contributors on the dedicated MIMIC git repository [29] or from the OMOP community forum [30]. Some trickier choices have been discussed in the MIMIC-OMOP git repository [31] and can be tracked in the commit logs.

The *implementation* was done through an extract-transform-load (ETL) process that consisted of Postgres scripts to extract information from the source or concept mapping tables and then transform it and load it into an OMOP target table. The scripts were managed sequentially through a main program. As a last resort, some modifications to the structural model of OMOP were made. A dedicated script recaps all of them and contains columns name modifications, new columns, column type modifications, or database indexing modifications. In particular, each source table has been given a unique global sequence incremented from 0, which serves as the primary key and links to the OMOP target tables. As a result, every record was uniquely identified, allowing us to chain the information with OMOP, while simplifying the maintenance of primary/foreign keys.

Although *evaluating* a structural model is difficult [32], several papers have attempted to assess the quality of the CDM [9,25]. The criteria developed by Khan et al [9], which refer to the Moody and Shanks metrics [32], were adapted to assess the quality of the data transformation (Table 1).

Table 1. Transformation quality evaluation metrics.

Data model dimension	Description
Completeness: structural mapping	Domain coverage: coverage of sources domains that are accommodated by the standard OMOP[a] model
Completeness: conceptual mapping	Data coverage: coverage of sources data concepts that mapped to the standard OMOP concept
Integrity	"Meaningful data relationships and constraints that uphold the intent of the data's original purpose" [9]
Flexibility	The ease to expand the standard model for new datatypes and concepts
Integration	The capacity of the standard model to use multiple terminologies and link them to standard ones
Implementability	The stability of the models, the community, and the cost of adoption
Understandability	The ease of the standard model to be understood
Simplicity	The ease of querying the standard model (the model should contain the minimum of concepts and relationship)

[a]OMOP: Observational Medical Outcomes Partnership.

In addition to the Moody and Shanks metrics, we provided a set of controls to guarantee correct transformation. To compare overall statistics, some SQL queries were set up to compare MIMIC and MIMIC-OMOP, and we provided basic characterizations of the populations. All tables were covered and tested through simple counts, aggregate counts, or distribution checks. We estimated the loss of information during the ETL process by measuring the percentage of both columns and rows lost in the process, as other previous studies have done [17]. It is important to note that we chose not to keep irrelevant information: for example, some rows are known to be invalid in MIMIC or some information is redundant. Each ETL script was tested using pgTAP, a unit testing framework for Postgres. Each unit test script checked whether a particular OMOP target table was correctly loaded. Integrity constraints (primary keys, foreign keys, nonnull columns) were included to apply integrity checks at ETL run time. The last part of the structural evaluation was Achilles software. It is open source analysis software produced by Observational Health Data Sciences and Informatics (OHDSI) [33]. Like many previous authors, we used Achilles to assess data quality [34]. This tool is used for data characterization, data quality assessment (Achilles Heel), and health observation data visualization. All the resulting tables are presented in the Results section.

Conceptual Mapping

The *conceptual mapping* aimed at aligning the MIMIC local terminologies to OMOP's standard ones. It consisted of 3 phases: integration, alignment, and evaluation.

The *integration* phase consisted of loading both types of terminologies into the OMOP vocabulary tables. The OMOP terminologies are provided by the Athena tool and were loaded with the associated programs. We used export with all terminologies without licensing limitations. The local terminologies were extracted from the multiple MIMIC tables and loaded into the OMOP CONCEPT table. When possible, relevant information from the original MIMIC tables was concatenated in the *concept_name* column. MIMIC local concepts were loaded with a *concept_id* identifier starting from 2 billion. In the OMOP CONCEPT table, MIMIC concepts could be distinguished with the *vocabulary_id* identifier equal to "MIMIC code" and a *domain_id* identifier targeting the OMOP table in which the corresponding data were stored. This domain information was used in the ETL to send the information

to the proper table. Following OMOP documentation, the conceptual mapping has to be performed before the structural mapping because the nature of the standard OMOP concepts guides in which table (domain) the information should be stored.

The *alignment* phase, aimed at standardizing local MIMIC codes into standard OMOP codes, had 4 distinct cases. In the first case, some MIMIC data were, by chance, already coded according to standard OMOP terminologies (eg, LOINC laboratory results) and, therefore, the standard and local concepts were the same. In the second case, MIMIC data were not coded according to the standard OMOP terminologies but the mapping was already provided by OMOP (eg, *International Classification of Diseases, 9th Revision* [ICD9]/Systematized Nomenclature of Medicine-Clinical Terms [SNOMED-CT]), so the domain tables were loaded accordingly. In the third case, terminology mapping was not provided, but it was small enough to be done manually in a few hours (eg, demographic status, signs, and symptoms). In the fourth case, terminology mapping was not provided and consisted of a large set of local terms (admission diagnosis, drugs). Next, only a subset of the most represented codes was manually mapped.

We chose to use simple SQL queries that were flexible enough to be queried on demand or to generate a prefilled csv with the best matches. We used Postgres full-text ranking features and linked local and standard candidates with a rating function based on their labels. This work was performed under the control of an intensivist.

The *evaluation* phase was both quantitative and qualitative. The quantitative evaluation measured the completeness of our work: the percentage of local concepts that were mapped to standard concepts. The qualitative evaluation assessed the correctness. For newly generated mappings, this consisted of manually tagging each mapping with a score between 0 and 1 and eventually writing a commentary on each mapped concept. In case where the mapping was provided by automatic OMOP terminology mapping, the evaluation was performed on a subset of concepts manually picked within each terminology.

Data Analytics

Beyond the model transformation and with regard to the OMOP standardization process, we performed some analysis. MIMIC provides a large number of SQL scripts for preprocessing and normalizing data, calculating derived scores, and defining cohorts. Some of them were implemented on top of the OMOP format to load the OMOP-derived tables.

A set of *general denormalized* tables was built on top of the original OMOP format and had the *concept_name* related to the *concept_id* columns. The CONCEPT table is a central element of OMOP, and therefore, it was involved in many joins to obtain the concept label. By precalculating the joins with the CONCEPT tables, the denormalized tables rendered faster calculation and simplified SQL queries.

In addition, a set of *specialized analytical tables* was built, in addition to the original OMOP tables. The MICROBIOLOGICALEVENTS table was a reorganization of the MEASUREMENT table data of microorganisms and associated susceptibility testing antibiotics. It was based on the MIMIC MICROBIOLOGICALEVENTS table. The ICUSTAYS table allowed us to quickly determine the patients admitted in resuscitation and was inspired by the MIMIC ICUSTAYS tables.

The OMOP NOTE_NLP table was originally designed to store the final or intermediate derived information and metadata from clinical notes. When definitive, the extracted information is intended to be moved to the dedicated domain or table and then reused as regular structured data. When the information is still intermediate, it is stored in the NOTE_NLP table and can be used for later analysis. To populate this table, we provided 2 information extraction pipelines. The first pipeline extracted numerical values, such as weight, height, body mass index, and left ventricular cardiac ejection fraction, from medical notes with a SQL script. The resulting structured numerical values were loaded into the measurement or observation tables according to their domain. The second pipeline *section extractor*, based on the Apache Unstructured Information Management Architecture (UIMA) framework, divided notes into sections to help analysts choose or avoid certain sections of their analysis. Section templates (eg, “Illness History”) were automatically extracted from text with regular expressions and then filtered to keep only the most frequent (frequency >1%).

A 48-hour open access datathon was set up in the Assistance Publique des Hopitaux de Paris (Paris AP-HP) in collaboration with the Massachusetts Institute of Technology (MIT), once the MIMIC-OMOP transformation was ready for research. This datathon was organized to evaluate OMOP as an alternative data model for accessing and analyzing MIMIC data during a real event. Scientific questions were prepared in an online forum where participants could introduce themselves and propose a topic or choose an existing one. OMOP was loaded into Apache HIVE 1.2.1 in ORC format. Users had access to the ORC data set from a web interface Jupyter Notebooks with Python, R, or Scala. A SQL web client allowed teams to write SQL queries from Presto to the same data set. The hadoop cluster was based on 5 computers with 16 cores and 220 GB of RAM. The MIMIC-OMOP data set was loaded from a Postgres instance to HIVE through Apache SQOOP 1.4.6 directly in ORC format. Participants also had access to the Schemaspy database physical model to access the OMOP physical data model with both table/column comments and key primary/foreign relationships materializing the relationships between the tables. All queries were logged.

Results

Data Transformation

All transformation processes are freely accessible to the public via the MIMIC-OMOP git repository maintained by MIT-LCP [8]. The git repository centralizes the various resources of this work, such as documentation, source code, unit tests, and questioning examples, discussions, and problem issues. It also indicates web resources, such as the physical data model for MIMIC and OMOP data sets and the Achilles Heel web client.

The MIMIC-OMOP conversion was performed by 2 developers (a data engineer and an intensivist) for 500 hours. This included ETL, git documentation, concept mapping, contributions, and

unit tests. ETL (with unit tests and generation of ready-to-load archive) on a subset of 100 patients lasted 5 minutes and enabled fast development cycles. ETL lasted 3 hours to process the whole MIMIC database. The resulting csv archive was almost the same size as the original archive, and MIMIC-OMOP was also the same size as MIMIC once loaded and indexed into Postgres.

Structural Mapping

The results of the structural mapping are presented in Table 2. Of the 37 OMOP tables, the ones related to hospital costs were not applicable, some tables related to derived data were not populated, and some tables related to vocabulary were preloaded with terminology information. The 26 tables of MIMIC were dispatched into 19 OMOP tables. The reduced number of tables resulted from the differences in the design of both models. OMOP stores all the terminologies in 1 table, whereas MIMIC has 1 table for each terminology. In addition, the same applies for facts data, which are grouped by nature in OMOP, while MIMIC tables are more specialized and respect the source EHR's design. For example, the MEASUREMENT table gathers measured information and combines 4 source tables, resulting in 365,181,104 rows, which is 20% more than the largest MIMIC table. To some extent, this is a regression in terms of performance.

Table 2. MIMIC[a]-OMOP[b] data flows.

OMOP tables	Number of rows (n)	MIMIC tables
CARE_SITE	93	transfers, service
COHORT_ATTRIBUTE	228,379	callout
CONCEPT	30,344	d_cpt, d_icd_procedures, d_items, d_labitems
CONDITION_OCCURRENCE	716,595	admissions, diagnosis_icd
DEATH	14,849	patients, admissions
DRUG_EXPOSURE	24,934,751	prescriptions, inputevents_cv, inputevents_mv
MEASUREMENT	365,181,104	chart/lab/microbiology/in/output events
NOTE	2,082,294	noteevents
NOTE_NLP	16,350,855	noteevents
OBSERVATION	6,721,040	admissions, chartevents, datetimevvents, drgcodes
OBSERVATION_PERIOD	58,976	patients, admissions
PERSON	46,520	patients, admissions
PROCEDURE_OCCURRENCE	1,063,525	cptevents, procedureevents_mv, procedure_icd
PROVIDER	7567	caregivers
SPECIMEN	39,874,171	chartevents, labevents, microbiologyevents
VISIT_OCCURRENCE	58,976	admissions
VISIT_DETAIL	271,808	admissions, transfers, service

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]OMOP: Observational Medical Outcomes Partnership.

Two important tables are provided by OMOP to model the relationship between the data: CONCEPT_RELATIONSHIP and FACT_RELATIONSHIP. We used them to bind the drugs into a solution for microbiology/antibiograms and for VISIT_DETAIL/CARE_SITE links. The following SQL query (Textbox 1) shows how a microorganism is linked to its susceptibility test by a FACT_RELATIONSHIP and illustrates the flexibility of the model. However, this flexibility affects the simplicity and the performance of the model by increasing the number of joins within SQL queries.

Textbox 1. Original table microbiology SQL query.

```
SELECT measurement_source_value
, value_as_concept_id
, concept_name
FROM measurement
JOIN concept resistance
ON value_as_concept_id = concept_id
JOIN fact_relationship
ON measurement_id = fact_id_2
JOIN
(
SELECT measurement_id AS id_is_staph
FROM measurement
WHERE
measurement_type_concept_id = 2000000007
-- 'Labs - Culture Organisms'
AND value_as_concept_id = 4149419
-- 'Staph aureus coag +'
AND measurement_concept_id = 46235217
-- 'Bacteria identified in Blood product
unit.autologous by Culture'
) staph ON id_is_staph = fact_id_1
WHERE TRUE
AND measurement_type_concept_id = 2000000008
-- 'Labs - Culture Sensitivity'
```

Table 3 presents the basic characterization of the MIMIC-OMOP population and assesses the overall quality of structural mapping. Fortunately, most statistics remain similar between the 2 versions, with few differences. Table 3 shows that MIMIC contains 61,532 intensive care stays, while OMOP contains 71,576 intensive care stays. This represents a 16% increase in stays.

Table 3. Baseline characteristics of MIMIC[a] versus OMOP[b].

Items	MIMIC	MIMIC-OMOP
Overall		
Persons (n)	46,520	46,520
Admissions (n)	58,976	58,976
ICU[c] stays (n)	71,575	61,532
Female gender, n (%)	20,399 (43.85)	20,399 (43.85)
Age (N=58,976)		
Mean	64 years, 4 months	64 years, 4 months
0-5 years, n (%)	8110 (13.75)	8110 (13.75)
6-15 years, n (%)	1 (0.001)	1 (0.001)
16-25 years, n (%)	1434 (2.43)	1434 (2.43)
26-45 years, n (%)	5962 (10.11)	5962 (10.11)
46-65 years, n (%)	17,375 (29.46)	17,375 (29.46)
66-80 years, n (%)	15,793 (26.78)	15,793 (26.78)
>80 years, n (%)	10,301 (17.47)	10,301 (17.47)
Other characteristics		
Emergency, n	42,071	42,071
Elective, n	7706	7706
Surgical patients, n	19,246	19,246
Length of hospital stay, days, median (Q1-Q3)	6.46 (3.74-11.79)	6.59 (3.84-11.88)
Length of ICU stay, days, median (Q1-Q3)	2.09 (1.10-4.48)	1.87 (0.95-3.87)
Mortality in ICU, n (%)	5814 (9)	5815 (9)
Mortality in hospital, n (%)	4511 (7)	4559 (6)
Lab measurements per admissions, mean	478	678
Procedures per admission, mean	4.6	4.6
Drugs per admission, mean	82.8	82.8
Exit diagnosis per admission, mean	11	11

[a]MIMIC: Medical Information Mart for Intensive Care.

[b]OMOP: Observational Medical Outcomes Partnership.

[c]ICU: intensive care unit.

By design, MIMIC aggregates information from various systems. Thus, the transfer information is divided into several tables, such as ADMISSIONS, TRANSFERS, and also ICUSTAYS, while OMOP centralizes this information in VISIT_DETAIL. We also added emergency stays as a normal location for patients throughout their hospital stay (unlike what had been done by MIMIC). The ICUSTAYS MIMIC table was not transformed, because it derives from the TRANSFER table and we decided to assign a new VISIT_DETAIL row for each ICU stay (based on the TRANSFER table), while MIMIC prefers to assign a new ICU stay if a new admission occurs more than 24 hours after the end of the previous stay. This table also showed an increase in the number of laboratory measurements per admission. This is because MIMIC-OMOP gathers laboratory data from both the MIMIC-dedicated LABORATORY table and the CHARTEVENTS table, which is usually not considered for this purpose. For laboratory tests, we put a specimen (ie, a blood sample) for many laboratory results (because 1 blood sample can be used for several tests), and we decided to create as many rows of samples as laboratory tests because the information was not present in MIMIC. The same was true when date information was not provided (start/end _datetime} for DRUG_EXPOSURE).

As mentioned in Table 4, 20%-80% of the source columns were not retained. Almost all were redundant or provided derived information. The main concern was the loss of some timestamps. For example, the MIMIC CHARTEVENTS table provides the *storetime* and *charttime* columns, but OMOP only provides 1 column to store timestamps. Thus, the MIMIC *storetime* column was eliminated during ETL, which was considered less valuable.

Table 4. Data lost.

Relationship	Rows lost, %	Columns lost, %
admissions	__[a]	30
callout	—	80
caregivers	—	50
chartevents	0.04	40
cptevents	—	60
datetimeevents	0.0001	50
diagnoses_icd	—	20
drugcodes	—	60
inputevents_cv	—	41
inputevents_mv	10.0	46
labevents	—	34
microbiologyevents	—	30
noteevents	0.04	19
outputevents	—	39
patients	—	50
prescriptions	—	16
procedureevents_mv	3.0	70
procedures_icd	—	40
services	—	34
transfers	—	47

[a]Not available.

As mentioned in the Methods section, incorrect entries were not kept in the process. Five MIMIC tables (INPUTEVENTS_MV, CHARTEVENTS, PROCEDUREEVENTS_MV, NOTEEVENTS, and DATETIMEEVENTS) had deleted rows in the ETL process. All of them were tagged in MIMIC as erroneous or cancelled.

A set of minor modifications of the OMOP table structure was made in order to fit the data. All character columns with limited length were modified to unlimited length since this could cause unpredictable truncation of content, while having no negative impact on the Postgres storage size or performance. The VISIT_OCCURRENCE and VISIT_DETAIL tables were corrected according to some discussions of the OHDSI forum. The NLP_NOTE table was extended with fields mentioned in online documentation but forgotten in the scripts. In addition, the *offset* column was divided into 2 integer-type columns because the offset term was a SQL reserved word and it made sense to fill the resulting *offset_begin* and *offset_end* columns with integer values.

All the PgTAP unit tests passed. Moreover, OMOP had a 100% match of the integrity constraints and the foreign key relationships of the data models. After 18 hours of computations, Achilles Heel issued 15 errors, 18 warnings, and 8 notifications. This result is good compared to other studies [27].

Conceptual Mapping

The results of the conceptual mapping's completeness are presented in Table 5. We have often mapped many source concepts to a unique standard *concept_id* because MIMIC provides a large number of equivalent concepts. For example, MIMIC provides 6 distinct concepts for body temperature: temperature C, temperature C (calc), temperature F, temperature F (calc), temperature Fahrenheit, and temperature Celsius. All of them were mapped to the LOINC “Body temperature”, and numerical values were normalized.

Table 5. Terminology mapping coverage.

OMOP[a] tables (domain)	Records, n	Mapped records, n (%)	Concept source, n	Mapped concepts source, n (%)
CARE_SITE	144	144 (100)	58	58 (100)
CONDITION_OCCURRENCE	716,595	644,936 (90)	6984	6565 (94)
DRUG_EXPOSURE	24,934,751	9,475,205 (38)	7398	4143 (56)
MEASUREMENT	40,141,521	29,303,310 (73)	1035	787 (76)
OBSERVATION	6,721,040	4,570,307 (68)	1440	1152 (80)
PERSON	93,040	93,040 (100)	43	43 (100)
PROCEDURE_OCCURRENCE	1,063,525	1,052,890 (99)	2203	2181 (99)
SPECIMEN	39,874,171	27,911,920 (70)	92	71 (77)
NOTE	2,082,294	2,082,294 (100)	15	15 (100)
VISIT_OCCURRENCE	176,928	176,928 (100)	34	34 (100)
VISIT_DETAIL	396,932	396,932 (100)	28	28 (100)

[a]OMOP: Observational Medical Outcomes Partnership.

OMOP's terminology coverage has already been rated as excellent [24]. We used the OMOP terminology mappings (National Drug Code [NDC]-RxNorm, ICD9-SNOMED, Common Procedural Terminology Fourth Revision [CPT4]-SNOMED) to standardize a consequent set of MIMIC nonstandard terminologies.

The automatic OMOP terminology mapping was evaluated by an intensivist. The results are in favor of good integration of the model. We checked 100 elements for each mapping used (NDC, ICD9, and CPT4). ICD9 and CPT4 were correctly mapped to SNOMED (100%). However, only 85% of NDCs were linked to a correct RxNorm code. This was partly due to an incorrect NDC drug code (from MIMIC) and partly because only 78% of NDC codes are mapped to RxNorm. Moreover, even if this does not seem to have affected our ETL, we know that some of ICD-9-CM codes can have a one-to-several match with SNOMED (28%) [35].

In several cases, OMOP had no suitable concepts for the ICU-specific cases. In particular, the VISIT_DETAIL table does not yet introduce relevant information and duplicate information from the VISIT_OCCURRENCE table. Therefore, we extended the concepts to track bed transfers and room transfers through *admitting_concept_id*, *discharge_to_concept_id*, or *visit_type_concept_id* columns. These added concepts were introduced with *concept_id* between 2 billion and 2.001 billion to distinguish them from OMOP concepts (0-2 billion) and MIMIC locals (>2.001 billion).

Some local concepts could not be mapped to standard ones. These unmapped concepts were linked with the *concept_id* = 0 and appeared in different cases. In the first case, the local concept has no equivalent in the standard vocabularies. In the second case, it has not yet been mapped and may have a standard equivalent. In the third case, the value is missing and cannot be mapped. In our opinion, although not all of these cases can be used for standard queries, they should have a different concept identifier in order to be treated differently (not just *concept_id* = 0). Some of the *domain_id* do not match the table name, and this makes sense because the OBSERVATION domain can be the MEASUREMENT table and vice versa. Although various types of information are stored in the MEASUREMENT table, the dedicated OMOP concepts for the *measurement_type_concept_id* column were not sufficient to distinguish them. Therefore, we added some *measurement_type* concepts (eg, Labs - Chemistry, Labs - Culture Organisms).

Analytics

Some MIMIC raw information was transformed and added to match the structural model. The laboratory textual values were split into operators, numeric values, and units, when needed, with a dedicated Postgres stored procedure. The free text conditions were normalized and mapped to standard OMOP codes to meet the conceptual model.

As indicated in the Methods section, we provided many *derived values*. Common derived information was introduced and loaded: corrected serum calcium, corrected serum potassium, the P/F ratio, corrected osmolarity, and the Simplified Acute Physiology Score (SAPS) II.

Denormalized derived tables improved SQL query performance and verbosity. In addition, the resulting tables were much more human-readable, with the concept label directly in the table and greatly reduced joins. Therefore, a little denormalization greatly improved the analysts' experience of the data model and simplicity by adding some redundancy in the data, while not interrupting existing SQL queries. Moreover, these normalized views were backward-compatible and remained standardized, allowing the creation of multicentric algorithms. We provided 2 examples of materialized specialized views derived from MICROBIOLOGYEVENTS and ICUSTAYS MIMIC that simplified the experience for scientists (Textbox 2). These results reflect the lack of simplicity of the model in its original form, but this can be easily overcome with such analytics tables. These results were in favor of good flexibility of the model, allowing us to store derived data.

Textbox 2. Optimized and denormalized microbiology table SQL query.

```
SELECT antibiotic_source_value,
antibiotic_interpretation_concept_id,
antibiotic_interpretation_concept_name
FROM microbiology
WHERE
organism_concept_id = 4149419
-- 'Staph aureus coag +'
AND specimen_concept_id = 46235217
-- 'Bacteria identified in Blood product
unit.autologous by Culture';
```

The note section *extraction pipeline* resulted in 1200 sections that were collected and then manually filtered to exclude false positives; 400 similar groups were highlighted. The extracted sections were not mapped to standard terminologies, such as the LOINC clinical document ontology (CDO). The reason for this is that the LOINC CDO decided not to keep these sections up to date, considering that they are not widely used [36].

The Paris AP-HP organized a datathon with MIMIC-OMOP, in which 160 participants from 25 teams had 48 hours to undertake a clinical project using the MIMIC-OMOP database. They launched around 15,000 queries, with a maximum duration of 1 minute. They got an opportunity to create mixed teams: clinicians brought the issues that required data mining, as well as their data expertise, while data scientists judged the technical feasibility and finally implemented the various analyses needed. Writing standard queries (ie, with standard concepts) requires knowing the organization of relational models (SQL) and also mastering the graphical nature of certain terminologies, such as SNOMED-CT, in order to capture all potential codes that might be related to the one analysts think of first. Overall, the teams quickly mastered the OMOP model and managed to produce results at the end of the datathon. These results were in favor of good understandibility and simplicity of the model.

Discussion

Principal Results

In this paper, we presented the transformation of the MIMIC database into the OMOP CDM and its evaluation. The first major contribution of this study is to provide a freely accessible data set in OMOP format that could be useful to researchers. The second major contribution is to share with the OMOP community some useful transformations dedicated to intensive care that can be reused on any OMOP data set. The last contribution is to evaluate the implementation of MIMIC into the OMOP CDM.

Lessons Learned

We observed that the OMOP CDM can be implemented at low cost and downstream of an existing architecture, since the scripts are freely available on the project's GitHub, for 8 different database management systems. The rationale of the data model can be understood through the numerous resources made available by the OHDSI community: tutorials, forums, working groups, and documentation. The structural mapping is carried out without difficulty as question marks can be raised with the community. The main difficulty remains the step of semantic mapping, especially in countries or institutions using local terminologies and vocabularies. Since the CDM model proposes to store both international and local vocabulary codes for each table, it is possible to start conducting studies using only the local codes. The mapping to the international codes can be carried out in a second phase, project by project, for the codes presented by each study. This will make it easier to spread out the difficulty of global mapping over thousands of codes.

Data Transformation

The choice of a simple SQL-based ETL over dedicated ETL software has several advantages. SQL, as a unique language, factors both people's knowledge and computer resources, allowing analysts to become implementers and revise code or contribute to transformations. SQL was also used for semantic mapping, and OHDSI provides Usagi [37]. The use of csv format for sharing information is simple and universal. Both SQL and CSV are standard and target a large community (physicians, engineers, and analysts) with translational profiles and is compatible with multiple technologies.

The calculation time of ETL on the Postgres instance on a modest personal computer is compatible with community work where the collaborator can clone the source code and configure a development instance to reproduce or improve the work.

Choosing a public GitHub repository for documentation and source code support allows analysts to learn more about the project and also learn how to contribute. The highly active OMOP forum is full of details and training. In contrast, the implementation guide suffers from not being as detailed and maintained. We believe that the OMOP community would greatly benefit from a systematic and concise synchronization between the forum, mailing lists, source code repository, and end-user documentation.

Any data transformation is likely to generate bugs that can later have an impact on medical research. The foundations of the relational database management system (RDBMS), such as transactions, standardization, and integrity constraints, are integrated safeguards that have been useful throughout the

process. In addition, the implemented unit tests ensure that past bugs are not repeated. An ideal but complex validation method would be to replicate existing MIMIC studies and ensure that the results are consistent across data models. The OHDSI Achilles tool completes our quality assessment. It is a surprisingly slow tool to process. The rules and their descriptions are difficult to understand. A more specific tool should be provided and described.

Another missing aspect is a set of quality tables for assessing and measuring data quality. MIMIC has a column to keep track of corrupted information. It would be interesting to be able to keep the disordered data in OMOP and enable research in the data cleaning/quality field. Although the OMOP-CDM provides rules to name columns, there are some mistakes, and we have to modify it. One the one hand, it is a problem for a CDM to contain errors, but on the other hand, it is easy to relay issues that are now corrected.

Data Analytics

It is important that OMOP maintain a level of standardization in order to simplify ETL and make it consistent. However, once done, it makes sense to give access to scientific data through more denormalized and specialized tables. There are many concerns about OMOP's performance and optimization. However, there will never be a perfect multipurpose case table, and it is the responsibility of data scientists to build their own, simplified, specialized tables for their research and to respond effectively and clearly to their needs.

The derived data integrate quite well into OMOP. We used the NOTE_NLP table to store information derived from notes, the MEASUREMENT table to store derived numerical information, and the COHORT_ATTRIBUTE table to store derived scores. However, it is not yet clear whether derived data should be stored by domain or whether they should be stored in dedicated derived tables. We found that there are no tables to track the source and description of these data.

The pipeline notes' section extractor we used was based on the Apache UIMA framework. Although some methods already exist to extract medical sections [38], the prior work of describing sections was too complex, and we opted for a naive approach.

Last but not least, as noted in the Introduction section, a good CDM for the ICU would allow for near real-time early warning systems and inference modeling on fresh data. OMOP is clearly designed to provide a static data set and does not have real-time ingestion and data versioning control mechanisms like EHRs usually do. Analysis of static data sets is essential for reproducible results. However, when the algorithms need to be moved to the bedside, it is necessary to have fresh data and a way of re-identifying the patient that OMOP does not yet provide. That said, a solution such as the HL7 FHIR is a great way to implement real-time inference from EHR data, and that is how the FHIR and OMOP are complementary. This has already been studied but needs further optimization [39].

The MIT regularly organizes datathons using their open-access databases [40-43]. From a human point of view, these events enable teamwork and collaboration between different specialties (ie, physicians, computer scientists, statisticians, data scientists), which can benefit from each other's expertise. This time, the datathon was also an opportunity for these profiles to collaborate, and it allowed novices to be introduced to the OMOP CDM and its analytical tools. The critical point in the conducting of such an event is related to the IT architecture, which must allow dozens of users to run large queries at the same time and to share scripts and results. We used a platform similar to the one used by Celi et al [41], with several analytical tools (Jupyter Notebook, Python, R, Scala).

The datathon showed that distributed platforms with basic hardware provide SQL tools for online analytical processing (OLAP) with excellent performance that overcomes RDBMS weaknesses. Therefore, OLAP takes advantage of SQL language analysis functions, such as grouping, windowing, assembling, and mathematical functions, that are often missing in NoSQL databases. Although some are open source, these distributed technologies are not easily accessible; however, cloud-based solutions are increasingly affordable for researchers.

The real-life test of the datathon revealed the strong need to make the physical data model accessible, including comments on columns and tables, and we discovered that an open source tool called schemaspy is helpful. In addition, we found that the GitHub repository is the best place to document and interact with the community.

The OMOP model is powerful because it allows a broad spectrum of analysis from specialized local models to evidence-based statistical analysis in an easy-to-learn and accessible format. The major complexity of this model is intrinsically linked to the terminologies' complexity with the use of its closure table [21].

Compared to the original MIMIC data model, working with OMOP offers the ability to write standard code and analyses that could benefit other international users.

The effectiveness of the OMOP model has some weaknesses because it seems to focus on consistency rather than performance. However, we have shown that it is easy to overcome these weaknesses and improve OMOP with design or technology optimization and a dedicated structure that ultimately remains a standard and is shareable because it derives from the original model.

Conclusions

The transformation of MIMIC into OMOP required efforts that remain reasonable. It is and always will be a work in progress because standard concept mapping is an almost infinite process with constant improvements. Fortunately, the published version of MIMIC-OMOP is search ready and already offers the same scope of data as the original MIMIC version and even more with the derived data. It is publicly available on the GitHub repository and have been designed to be easily revised, copied, or enriched according to the OMOP or MIMIC philosophy by any users who know SQL.

Acknowledgments

We acknowledge the Massachusetts Institute of Technology and the Observational Health Data Sciences and Informatics community for their support.

Conflicts of Interest

None declared.

References

1. Angus DC, Kelley MA, Schmitz RJ, White A, Popovich J, Committee on Manpower for Pulmonary and Critical Care Societies (COMPACCS). Caring for the critically ill patient. Current and projected workforce requirements for care of the critically ill and patients with pulmonary disease: can we meet the requirements of an aging population? JAMA 2000 Dec 06;284(21):2762-2770. [doi: 10.1001/jama.284.21.2762] [Medline: 11105183]
2. Azoulay E, Alberti C, Legendre I, Buisson CB, Le Gall JR, European Sepsis Group. Post-ICU mortality in critically ill infected patients: an international study. Intensive Care Med 2005 Jan;31(1):56-63. [doi: 10.1007/s00134-004-2484-1] [Medline: 15526186]
3. Vincent J. Is the current management of severe sepsis and septic shock really evidence based? PLoS Med 2006 Sep;3(9):e346 [FREE Full text] [doi: 10.1371/journal.pmed.0030346] [Medline: 16933970]
4. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform 2014 Aug 15;9:97-104 [FREE Full text] [doi: 10.15265/IY-2014-0003] [Medline: 25123728]
5. Zhang Y, Guo S, Han L, Li T. Application and exploration of big data mining in clinical medicine. Chin Med J (Engl) 2016 Mar 20;129(6):731-738 [FREE Full text] [doi: 10.4103/0366-6999.178019] [Medline: 26960378]
6. Safran C. Reuse of clinical data. Yearb Med Inform 2014 Aug 15;9:52-54 [FREE Full text] [doi: 10.15265/IY-2014-0013] [Medline: 25123722]
7. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52 [FREE Full text] [doi: 10.15265/IY-2017-007] [Medline: 28480475]
8. 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]
9. Kahn MG, Batson D, Schilling LM. Data model considerations for clinical effectiveness researchers. Med Care 2012 Jul;50 Suppl:S60-S67 [FREE Full text] [doi: 10.1097/MLR.0b013e318259bff4] [Medline: 22692260]
10. Gagne JJ. Common models, different approaches. Drug Saf 2015 Aug;38(8):683-686. [doi: 10.1007/s40264-015-0313-9] [Medline: 26088718]
11. Platt R, Lieu T. Data enclaves for sharing information derived from clinical and administrative data. JAMA 2018 Aug 28;320(8):753-754 [FREE Full text] [doi: 10.1001/jama.2018.9342] [Medline: 30083726]
12. Xu Y, Zhou X, Suehs BT, Hartzema AG, Kahn MG, Moride Y, et al. A comparative assessment of observational medical outcomes partnership and mini-sentinel common data models and analytics: implications for active drug safety surveillance. Drug Saf 2015 Aug;38(8):749-765. [doi: 10.1007/s40264-015-0297-5] [Medline: 26055920]
13. Madigan D, Ryan PB, Schuemie M, Stang PE, Overhage JM, Hartzema AG, et al. Evaluating the impact of database heterogeneity on observational study results. Am J Epidemiol 2013 Aug 15;178(4):645-651 [FREE Full text] [doi: 10.1093/aje/kwt010] [Medline: 23648805]
14. Morgenstern H, Rafaely B. Spatial reverberation and dereverberation using an acoustic multiple-input multiple-output system. J Audio Eng Soc 2017 Feb 17;65(1/2):42-55. [doi: 10.17743/jaes.2016.0063]
15. Klungel O, Kurz X, de Groot MCH, Schlienger R, Tcherny-Lessenot S, Grimaldi L, et al. Multi-centre, multi-database studies with common protocols: lessons learnt from the IMI PROTECT project. Pharmacoepidemiol Drug Saf 2016 Mar;25 Suppl 1:156-165 [FREE Full text] [doi: 10.1002/pds.3968] [Medline: 27038361]
16. 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]
17. FitzHenry F, Resnic FS, Robbins SL, Denton J, Nookala L, Meeker D, et al. Creating a common data model for comparative effectiveness with the observational medical outcomes partnership. Appl Clin Inform 2015;6(3):536-547 [FREE Full text] [doi: 10.4338/ACI-2014-12-CR-0121] [Medline: 26448797]
18. Lamer A, Depas N, Doutreligne M, Parrot A, Verloop D, Defebvre M, et al. Transforming French electronic health records into the Observational Medical Outcome Partnership's common data model: a feasibility study. Appl Clin Inform 2020 Jan;11(1):13-22 [FREE Full text] [doi: 10.1055/s-0039-3402754] [Medline: 31914471]

19. 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]
20. Observational Health Data Sciences and Informatics. OHDSI: Observational Health Data Sciences and Informatics. 2021 Dec 01. URL: https://www.ohdsi.org/ [accessed 2019-05-03]
21. Karwin B. Keeping It Simple: Rendering Trees with Closure Tables. URL: https://karwin.blogspot.com/2010/03/rendering-trees-with-closure-tables.html [accessed 2021-11-01]
22. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19(1):54-60 [FREE Full text] [doi: 10.1136/amiajnl-2011-000376] [Medline: 22037893]
23. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012 Aug;45(4):689-696 [FREE Full text] [doi: 10.1016/j.jbi.2012.05.002] [Medline: 22683994]
24. Shamsuzzoha B, Vojtech H, Joydeep G. Conversion of MIMIC to OHDSI CDM. URL: https://www.ohdsi.org/web/wiki/lib/exe/fetch.php?media=resources:2016_ohdsi_paper_mimic_bayzid.pdf [accessed 2021-12-01]
25. Garza M, Del Fiol G, Tenenbaum J, Walden A, Zozus MN. Evaluating common data models for use with a longitudinal community registry. J Biomed Inform 2016 Dec;64:333-341 [FREE Full text] [doi: 10.1016/j.jbi.2016.10.016] [Medline: 27989817]
26. Chronaki C, Shahin A, Mark R. Designing reliable cohorts of cardiac patients across MIMIC and eICU. Comput Cardiol (2010) 2015;42:189-192 [FREE Full text] [doi: 10.1109/CIC.2015.7408618] [Medline: 27774488]
27. HL7. Welcome to FHIR®. URL: https://www.hl7.org/fhir/ [accessed 2021-12-01]
28. Athena. URL: http://athena.ohdsi.org/search-terms/terms [accessed 2019-01-03]
29. MIT Laboratory for Computational Physiology. MIT-LCP/mimic-code. URL: https://github.com/MIT-LCP/mimic-code [accessed 2021-12-01]
30. Observational Health Data Sciences and Informatics. OHDSI Forums. URL: https://forums.ohdsi.org/ [accessed 2021-12-01]
31. MIT Laboratory for Computational Physiology. MIT-LCP/mimic-omop. URL: https://github.com/MIT-LCP/mimic-omop [accessed 2021-12-01]
32. Moody DL, Shanks GG. Improving the quality of data models: empirical validation of a quality management framework. Inf Syst 2003 Sep;28(6):619-650. [doi: 10.1016/s0306-4379(02)00043-1]
33. Observational Health Data Sciences and Informatics. OHDSI/Achilles. URL: https://github.com/OHDSI/Achilles [accessed 2019-10-17]
34. 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]
35. U.S. National Library of Medicine. ICD-9-CM Diagnostic Codes to SNOMED CT Map Internet. URL: https://www.nlm.nih.gov/research/umls/mapping_projects/icd9cm_to_snomedct.html [accessed 2021-12-01]
36. Logical Observation Identifiers Names and Codes. LOINC Version 2.63 and RELMA Version 6.22 Are Now Available. URL: https://loinc.org/news/loinc-version-2-63-and-relma-version-6-22-are-now-available/ [accessed 2021-12-01]
37. Observational Health Data Sciences and Informatics. OHDSI/Usagi. URL: https://github.com/OHDSI/Usagi [accessed 2021-12-01]
38. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. J Am Med Inform Assoc 2009;16(6):806-815 [FREE Full text] [doi: 10.1197/jamia.M3037] [Medline: 19717800]
39. OMOPonFHIR. The FHIR Project at Georgia Tech Internet. URL: http://omoponfhir.org/ [accessed 2021-12-01]
40. Aboab J, Celi LA, Charlton P, Feng M, Ghassemi M, Marshall DC, et al. A "datathon" model to support cross-disciplinary collaboration. Sci Transl Med 2016 Apr 06;8(333):333ps8 [FREE Full text] [doi: 10.1126/scitranslmed.aad9072] [Medline: 27053770]
41. Celi LA, Lokhandwala S, Montgomery R, Moses C, Naumann T, Pollard T, et al. Datathons and software to promote reproducible research. J Med Internet Res 2016 Aug 24;18(8):e230 [FREE Full text] [doi: 10.2196/jmir.6365] [Medline: 27558834]
42. Luo EM, Newman S, Amat M, Charpignon M, Duralde ER, Jain S, et al. MIT COVID-19 datathon: data without boundaries. BMJ Innov 2021 Jan 31;7(1):231-234 [FREE Full text] [doi: 10.1136/bmjinnov-2020-000492] [Medline: 33437494]
43. Li P, Xie C, Pollard T, Johnson AEW, Cao D, Kang H, et al. Promoting secondary analysis of electronic medical records in China: summary of the PLAGH-MIT Critical Data Conference and Health Datathon. JMIR Med Inform 2017 Nov 14;5(4):e43 [FREE Full text] [doi: 10.2196/medinform.7380] [Medline: 29138126]

Abbreviations

CDM: common data model

CDO: clinical document ontology
CPT4: Common Procedural Terminology Fourth Revision
EHR: electronic health record
ETL: extract-transform-load
FHIR: Fast Healthcare Interoperability Resources
ICD9: International Classification of Diseases, 9th Revision
ICU: intensive care unit
LOINC: Logical Observation Identifiers Names and Codes
MIMIC: Medical Information Mart for Intensive Care
MIT: Massachusetts Institute of Technology
NDC: National Drug Code
OHDSI: Observational Health Data Sciences and Informatics
OLAP: online analytical processing
OMOP: Observational Medical Outcomes Partnership
RDBMS: relational database management system
SAPS: Simplified Acute Physiology Score
SNOMED: Systematized Nomenclature of Medicine
SNOMED-CT: Systematized Nomenclature of Medicine-Clinical Terms
UIMA: Unstructured Information Management Architecture

Edited by C Lovis; submitted 04.06.21; peer-reviewed by SD Boie, S Wei; comments to author 23.09.21; revised version received 03.10.21; accepted 05.10.21; published 14.12.21.

Please cite as:
Paris N, Lamer A, Parrot A
Transformation and Evaluation of the MIMIC Database in the OMOP Common Data Model: Development and Usability Study
JMIR Med Inform 2021;9(12):e30970
URL: https://medinform.jmir.org/2021/12/e30970
doi:10.2196/30970
PMID:34904958

Original Paper

Deep Learning–Assisted Burn Wound Diagnosis: Diagnostic Model Development Study

Che Wei Chang[1,2], MD; Feipei Lai[1], PhD; Mesakh Christian[3], BSc; Yu Chun Chen[3], BSc; Ching Hsu[1], MSc; Yo Shen Chen[2], MD, PhD; Dun Hao Chang[2,4], MD; Tyng Luen Roan[2], MD; Yen Che Yu[2], MD

[1]Graduate Institute of Biomedical Electronics & Bioinformatics, National Taiwan University, Taipei, Taiwan

[2]Division of Plastic and Reconstructive Surgery, Department of Surgery, Far Eastern Memorial Hospital, New Taipei, Taiwan

[3]Department of Computer Science & Information Engineering, National Taiwan University, Taipei, Taiwan

[4]Department of Information Management, Yuan Ze University, Chung-Li, Taiwan

Corresponding Author:
Feipei Lai, PhD
Graduate Institute of Biomedical Electronics & Bioinformatics
National Taiwan University
Room 419, Computer Science and Information Engineering-Der Tian Hall, No 1
Roosevelt Road, Sec 4
Taipei, 106319
Taiwan
Phone: 886 2 3366 4888 ext 419
Email: flai@ntu.edu.tw

Abstract

Background: Accurate assessment of the percentage total body surface area (%TBSA) of burn wounds is crucial in the management of burn patients. The resuscitation fluid and nutritional needs of burn patients, their need for intensive unit care, and probability of mortality are all directly related to %TBSA. It is difficult to estimate a burn area of irregular shape by inspection. Many articles have reported discrepancies in estimating %TBSA by different doctors.

Objective: We propose a method, based on deep learning, for burn wound detection, segmentation, and calculation of %TBSA on a pixel-to-pixel basis.

Methods: A 2-step procedure was used to convert burn wound diagnosis into %TBSA. In the first step, images of burn wounds were collected from medical records and labeled by burn surgeons, and the data set was then input into 2 deep learning architectures, U-Net and Mask R-CNN, each configured with 2 different backbones, to segment the burn wounds. In the second step, we collected and labeled images of hands to create another data set, which was also input into U-Net and Mask R-CNN to segment the hands. The %TBSA of burn wounds was then calculated by comparing the pixels of mask areas on images of the burn wound and hand of the same patient according to the rule of hand, which states that one's hand accounts for 0.8% of TBSA.

Results: A total of 2591 images of burn wounds were collected and labeled to form the burn wound data set. The data set was randomly split into training, validation, and testing sets in a ratio of 8:1:1. Four hundred images of volar hands were collected and labeled to form the hand data set, which was also split into 3 sets using the same method. For the images of burn wounds, Mask R-CNN with ResNet101 had the best segmentation result with a Dice coefficient (DC) of 0.9496, while U-Net with ResNet101 had a DC of 0.8545. For the hand images, U-Net and Mask R-CNN had similar performance with DC values of 0.9920 and 0.9910, respectively. Lastly, we conducted a test diagnosis in a burn patient. Mask R-CNN with ResNet101 had on average less deviation (0.115% TBSA) from the ground truth than burn surgeons.

Conclusions: This is one of the first studies to diagnose all depths of burn wounds and convert the segmentation results into %TBSA using different deep learning models. We aimed to assist medical staff in estimating burn size more accurately, thereby helping to provide precise care to burn victims.

(JMIR Med Inform 2021;9(12):e22798) doi:10.2196/22798

KEYWORDS

deep learning; semantic segmentation; instance segmentation; burn wounds; percentage total body surface area

Introduction

Background

According to the World Health Organization, an estimated 265,000 deaths occur each year from burn injuries. In the United States, burn injuries result in 10 million visits to the emergency department and 40,000 patients requiring hospitalization annually. The most critical aspect of managing burn injuries is the accurate calculation of the burn area, expressed as percentage total body surface area (%TBSA). However, many articles have reported discrepancies in the %TBSA diagnosed by different doctors. In adult burn injuries, Harish et al reported that overestimation by the referring institution occurred in 53% of cases and that the difference was statistically significant [1]. In child burn injuries from a national survey, Baartmans et al reported that burn size was often overestimated by referrers, by up to 30% TBSA, while underestimation was up to 13% TBSA [2].

There are 2 types of inaccurate estimations of burn injuries: misdiagnosis of burn depth and miscalculation of burn area. Misdiagnosis of burn depth comes from the dynamic nature of wound change. The initial presentation of burn depth may be quite different from the presentation several days after injury. Hence, the reported accuracy of diagnosis of burn depth is only 64% to 76% among experienced burn surgeons [3]. When evaluations are performed by less experienced practitioners, the accuracy declines to 50%. Fortunately, many technologies have been developed for accurate diagnosis of burn depth, such as laser Doppler imaging (LDI), infrared thermography, and photoacoustic imaging [4-7]. For example, LDI, which is based on perfusion in the burn area, provides information that is highly correlated with burn wound healing potential. Healing potential is a practical indicator of burn depth.

Though the assessment of burn depth with such technologies is often satisfactory, miscalculation of burn area may be hard to avoid. Such miscalculation often occurs when an area of irregular shape is estimated by comparing it with another area of irregular shape, for example, estimating the %TBSA of an irregularly shaped burn area on the upper extremity of an adult using the estimation that the upper extremity has roughly 7% to 9% TBSA as a guide [8,9]. In an interesting study, Parvizi et al reported that even when participants reached consensus on the margin of the burn wound, their estimations of %TBSA were still different [10]. The difference in %TBSA resulted in discrepancies in estimating the amount of resuscitation fluid needed by as much as 5280 mL using the Parkland formula. Clearly, there is an unmet need to improve the accuracy of burn diagnosis.

Machine learning has many applications in the field of medicine, such as in drug development and disease diagnosis [11-14]. Although machine learning has also been implemented in many aspects of surgery, its application in burn care is relatively rare [15,16]. Burn care is a field where human error can be reduced by computer assistance.

Prior Work

Early work in the use of machine learning to assist burn diagnosis focused on classification of burn depth (Table 1). Since burn injuries result in a mixture of different burn depths, most images of burn wounds cannot be simply classified as superficial partial burn, deep partial burn, or full thickness burn. Before images of burn wounds are input for feature extraction, the images need to be processed.

Table 1. Segmentation of burn wounds.

Study	Image database	Model	Performance metric	Objective
Serrano et al [17]	38 images	Fuzzy-ARTMAP	Accuracy 88.57%	Burn depth
Acha et al [18]	50 images	Fuzzy-ARTMAP	Accuracy 82.26%	Burn depth
Acha et al [19]	50 images	SVM[a], Fuzzy-ARTMAP	Error rate 0.7%	Burn depth
Acha et al [20]	74 images	KNN[b], MDS[c]	Accuracy 83.8%	Need for skin grafts
Serrano et al [21]	94 images	SVM, MDS	Accuracy 79.73%	Need for skin grafts
Cirillo et al [22]	23 images	VGG16, GoogleNet, ResNet50, ResNet101	Accuracy 90.54%	Burn depth
Despo et al [23]	749 images	AlexNet, VGG16, GoogleNet	Accuracy 85%	Burn area segmentation, burn depth
Jiao et al [24]	1000 images	Mask R-CNN	DC[d] 84.51%	Burn area segmentation
Our study	2591 images	Mask R-CNN, U-Net	DC 94%	Estimation of burn %TBSA[e]

[a]SVM: support vector machine.

[b]KNN: K-nearest neighbor.

[c]MDS: multidimensional scaling.

[d]DC: Dice coefficient.

[e]%TBSA: percentage total body surface area.

Small Regions of Images

The most common method of addressing different burn depths in a given image is to select small regions of the image, called boxes, for processing. These small boxes are then transformed into a red/green/blue (RGB) matrix in a color coordinate system. The relative distance of each of the pixels from the others is then calculated and a threshold is set to check whether the box is homogeneous in texture and color. Homogeneous boxes are classified into different burn depths and input for machine learning.

Acha and Serrano collected 62 images of burn wounds with a resolution of 1536×1024 pixels. They selected regions of only 49×49 pixels from the images and classified these small boxes into 5 appearances to yield 250 images. They input the data set into Fuzzy-ARTMAP for training. A neural network was then used to classify burns into the 3 aforementioned types of burn depths with a success rate of 82% to 88% [17,18]. Later, they reduced the error rate from 1.6% to 0.7% by applying 5-fold cross-validation to the data sets and used support vector machine (SVM) to perform the classification [19]. In 2 subsequent studies, they further applied multidimensional scaling combining SVM and k-nearest neighbor classification to predict the need for a skin graft, with success rates of 79.73% and 83.8%, respectively [20,21].

Continuous Monitoring

Another method used to get the burn depths of a region corresponding to any specified pixels of the images of a burn wound is to record the wound from the time of injury to complete healing with the same protocol. Cirillo et al continuously collected images from the same burn wound until it healed [22]. They were then able to draw lines on the image corresponding to healing time and divide the area into 4 types of burn depths. To be more precise, they used the method mentioned above to extract small regions of the images (676 regions of 224×224 pixels from 23 images of 3456×2304 pixels). They then input these square regions of interest (RoIs) into several pretrained convolutional neural network (CNN) models, such as VGG19, ResNet18, ResNet50, and ResNet101. ResNet101 showed the best classification results with an average accuracy of 0.8166.

Goal of This Study

The use of machine learning in burn diagnosis to classify burn depth is currently quite limited. Technologies, such as LDI and thermography, are readily available and far more commonly employed. The treatment of burn injury may last for days or months. Without the use of special technologies, burn depth can still be determined by clinical assessment during the course of treatment. Recently, CNNs have been used in burn diagnosis to segment burn wounds. Despo et al reported a mean intersection over union (IoU) of around 0.7 with a fully convolutional network (FCN) [23]. Jiao et al reported a mean Dice coefficient (DC) of 0.85 with Mask R-CNN [24]. Such segmentation results could further be used to calculate %TBSA. This is important because all formulae for emergent fluid resuscitation (eg, the Parkland formula = %TBSA × body weight × 4) and calorie needs (eg, the Curreri formula = 25 × body weight + 40 × %TBSA) are based on %TBSA.

In this study, we implemented deep learning models to segment burn wounds and perform conversion to %TBSA based on the number of pixels. We tried to decrease the human error of estimating an area of irregular shape by inspection. We aimed to help medical staff obtain accurate formulae to aid in making decisions about triage, acute management, and transfer of burn patients.

Methods

Image Acquisition

This study was approved by the research ethics review committee of Far Eastern Hospital (number 109037-F). We reviewed the medical records of patients in Far Eastern Hospital from January 2016 to December 2019 with ICD9 codes 940-948, 983, and 994. We collected the images of burn wounds from their medical records and saved them as JPG files. These images were assigned random numbers for deidentification and were randomly presented to 2 out of 5 burn surgeons for labeling.

Labeling and Processing

Since many burn wounds have a mixture of different burn depths, the images were roughly classified into the following 3 categories: superficial/superficial partial burn, deep partial burn, and full thickness burn. Clinically, the color of superficial/superficial partial burns is red or pink, and the color of deep partial burns is dark pink to blotchy red. Blistering is common in superficial partial burns and is also present in deep partial burns of a relatively large size. Full thickness burns are white, waxy, or charred without blisters. All images were co-labeled by 2 burn surgeons to yield a single consensus result. The margins of the burn wounds were labeled without regard to burn depth with the labeling tool *LabelMe* and saved as JSON files. A burn wound image was excluded if the wound was on the face; it involved tattooed skin; it was coated with burn ointment; it appeared to have undergone an intervention, such as debridement or skin graft; or no agreement was reached on the margin of the burn wound by the 2 burn surgeons.

Since the images of burn wounds were collected from various medical records, their sizes were not uniform and ranged from 4000×3000 to 2736×1824 to 2592×1944 pixels. All labeled images were resized to 512×512 pixels. The data set of burn wounds was randomly split in a ratio of 8:1:1 into 3 sets for training, validation, and testing. We applied 2 deep learning architectures, U-Net and Mask R-CNN, in combination with 2 different backbones, ResNet50 and ResNet101, to segment these images.

Evaluation Metrics

The DC and IoU are 2 common metrics used to assess segmentation performance, whereas precision, recall, and accuracy are common metrics for assessing classification performance. The DC is twice the area of the intersection of the ground truth and prediction divided by the sum of their areas. It is given as follows:

where TP (true positive) denotes the number of correctly classified burn pixels, FP (false positive) denotes the number of mistakenly classified burn pixels, and FN (false negative) denotes the number of mistakenly classified nonburn pixels.

The IoU denotes the area of the intersection of the ground truth and prediction divided by the area of their union. It is given as follows:

Precision is defined as the ratio of burn pixels that models correctly classified in all predicted pixels. It is also called positive predictive value and is given as follows:

$$\text{Precision} = TP / (TP + FP) \quad \textbf{(3)}$$

Recall is defined as the ratio of burn pixels that are correctly classified in all actual burn pixels. It is also called sensitivity and is given as follows:

$$\text{Recall} = TP / (TP + FN) \quad \textbf{(4)}$$

Accuracy denotes the percentage of correctly classified pixels. It is given as follows:

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN) \quad \textbf{(5)}$$

where TN (true negative) denotes the number of correctly classified nonburn pixels.

Semantic Segmentation: U-Net

The convolutions in the U-Net path can be replaced with a deep network framework, such as the ResNet framework, which can explore and learn more features from the data (Multimedia Appendix 1). Then, the networks can be initialized using pretrained model weights derived from large-scale object detection, segmentation, and captioning data sets such as ImageNet and COCO. In our case, we trained our model using 2 different backbones, ResNet101 and ResNet50, with weights from the pretrained ImageNet model (Table 2). The standard augmentations of images we used were rotations, shifts, scale, gaussian blur, and contrast normalization. The standard Dice loss was chosen as the loss function. The formula is as follows:

The term is used to avoid the issue of dividing by 0 when precision and recall are empty.

Table 2. Configuration of the models.

Variable	Mask R-CNN	U-Net
Number of classes	1	1
Backbone	ResNet101 & ResNet50	ResNet101 & ResNet50
Regional proposal network anchor scales	8, 16, 32, 64, 128	N/A[a]
Train RoIs[b] per image, n	128	N/A
Anchors per image, n	256	N/A
Learning rate	0.0001 (initial rate, change in different epochs)	0.001
Learning momentum	0.9	0.9
Weight decay	0.0001	N/A
Batch size	8	8
Image dimensions	512×512	512×512

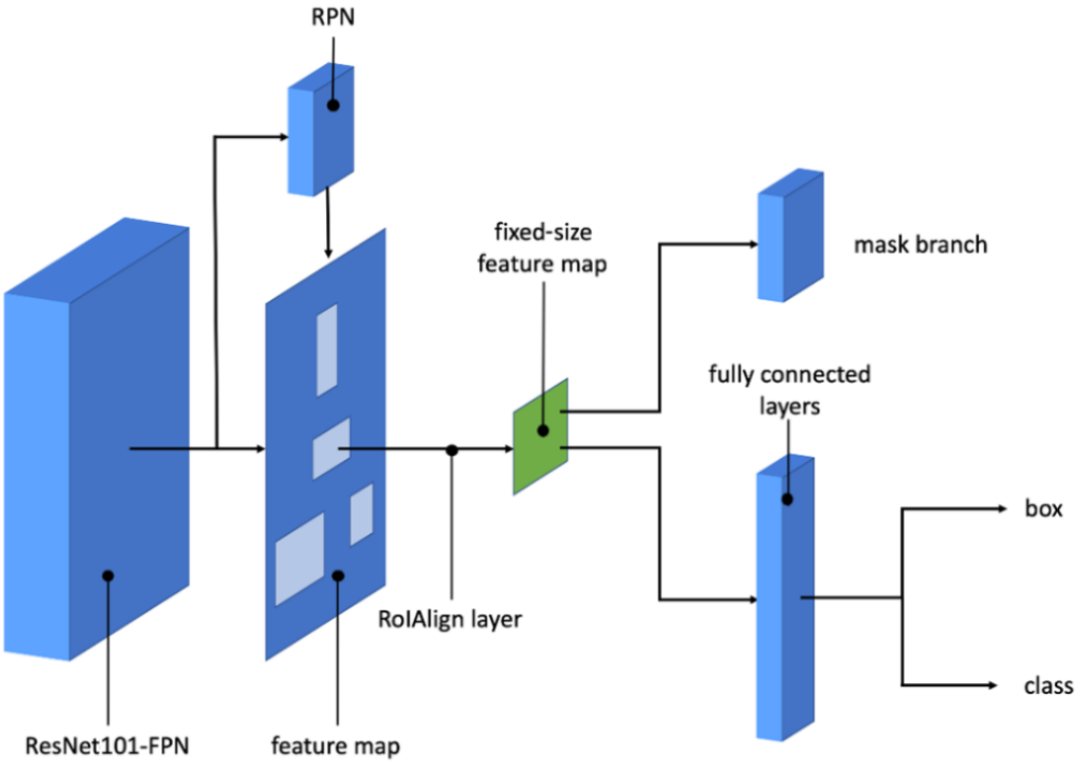
[a]N/A: not applicable.

[b]RoI: region of interest.

Instance Segmentation: Mask R-CNN

In our implementation of Mask R-CNN, we trained our model using ResNet101 and ResNet50 with weights from the pretrained COCO model (Table 1). Mask R-CNN uses a multitask loss function given by $L = L_{class} + L_{box} + L_{mask}$ (Figure 1). The L_{class} component contains the regional proposal network (RPN) class loss (failure of the RPN to separate object prediction from background) added to the Mask R-CNN class loss (failure of Mask R-CNN object classification). The L_{box} component contains the RPN bounding box loss (failure of object localization or bounding by the RPN) added to the Mask R-CNN bounding box loss (failure of object localization or bounding by Mask R-CNN). The last component L_{mask} loss constitutes the failure of Mask R-CNN object mask segmentation.

Figure 1. Mask R-CNN architecture with ResNet101. FPN: feature pyramid network; RoI: region of interest; RPN: regional proposal network.



Burn Segmentation to %TBSA

When the burn wounds are correctly segmented, the final step is to convert the pixels to %TBSA. To solve this problem, we applied the rule of hand/palm. The original rule is that a person's hand with digits accounts for 1% TBSA. It is the most common method of estimating burn %TBSA [25,26]. Recent studies have shown that a hand without digits represents precisely 0.5% TBSA (the rule of palm) and a hand with digits should be adjusted to around 0.8% TBSA (the rule of hand) [8]. If we use deep learning models to segment a patient's burn wounds as well as hands, we can then convert the segmentation result of burn wounds into %TBSA.

To produce the data set of hands and the data set of palms, we collected images of both volar hands from our colleagues. For each image, we labeled the hand with digits and without digits corresponding to the rule of hand and the rule of palm, respectively. These 2 data sets were split in a ratio of 8:1:1 into training, validation, and testing sets as well. The hand data set and the palm data set were processed according to the previous methods for burn wounds. The %TBSA of a burn wound can be calculated by comparing the mask area of the burn wound with the mask area of the hand or palm of the same patient. The formula is given by:

where M_{burn} is the number of pixels of the masked burn area, M_{hand} is the number of pixels of the masked hand area (0.8% TBSA), M_{palm} is the number of pixels of the masked palm area (0.5% TBSA), D_{burn} is the filming distance of the image of the patient's burn wound, and D_{hand} is the filming distance of the image of the patient's hand.

Results

Segmentation of Burn Wounds

There were 3 data sets used in our study, 1 each for burn wounds, hands, and palms. For the burn wound data, we collected 3571 images from the medical records of Far Eastern Hospital, 980 of which were excluded (mostly because the burn wounds had undergone interventions, and some because they were coated with burn ointment). The 2591 selected images were labeled and included in the burn wound data set. Among these images, 2073 were used as the training set and 259 were used as the validation set. The remaining 259 images were preserved as the testing set.

In our study, there was only 1 class in the ground truth. From the definitions of the DC and IoU, they have the relation of $1/2 \times DC \leq IoU \leq DC$ and perfect positive correlation. We used DC as our main metric to evaluate segmentation performance because it penalizes false negatives more than IoU does, and it is better to overestimate burn size than underestimate it.

Both U-Net and Mask R-CNN had better segmentation performance with the ResNet101 backbone than with ResNet50 (Table 3 and Table 4). The improvement was obvious in U-Net (DC: 0.8545 vs 0.8077) but negligible in Mask R-CNN (DC: 0.9496 vs 0.9493). Under the same backbone, Mask R-CNN had better performance in burn wound segmentation and classification than U-Net. Mask R-CNN with ResNet101 had the best segmentation result with a DC of 0.9496.

Figures 2-4 illustrate the performance of the 2 models in segmenting different burn depths. Both Mask R-CNN and U-Net showed poor segmentation results when they encountered small scattered burns (Figure 5).

Table 3. Segmentation results of burn wounds with ResNet101.

Variable	U-Net	Mask R-CNN
Mean DC[a]	0.8545	0.9496
Mean IoU[b]	0.7782	0.9089
Mean precision	0.9041	0.9613
Mean recall	0.8541	0.9390
Mean accuracy	0.7893	0.9130

[a]DC: Dice coefficient.

[b]IoU: intersection over union.

Table 4. Segmentation results of burn wounds with ResNet50.

Variable	U-Net	Mask R-CNN
Mean DC[a]	0.8077	0.9493
Mean IoU[b]	0.7190	0.9075
Mean precision	0.8947	0.9610
Mean recall	0.8002	0.9382
Mean accuracy	0.7331	0.9117

[a]DC: Dice coefficient.

[b]IoU: intersection over union.

Figure 2. Superficial partial burn. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.

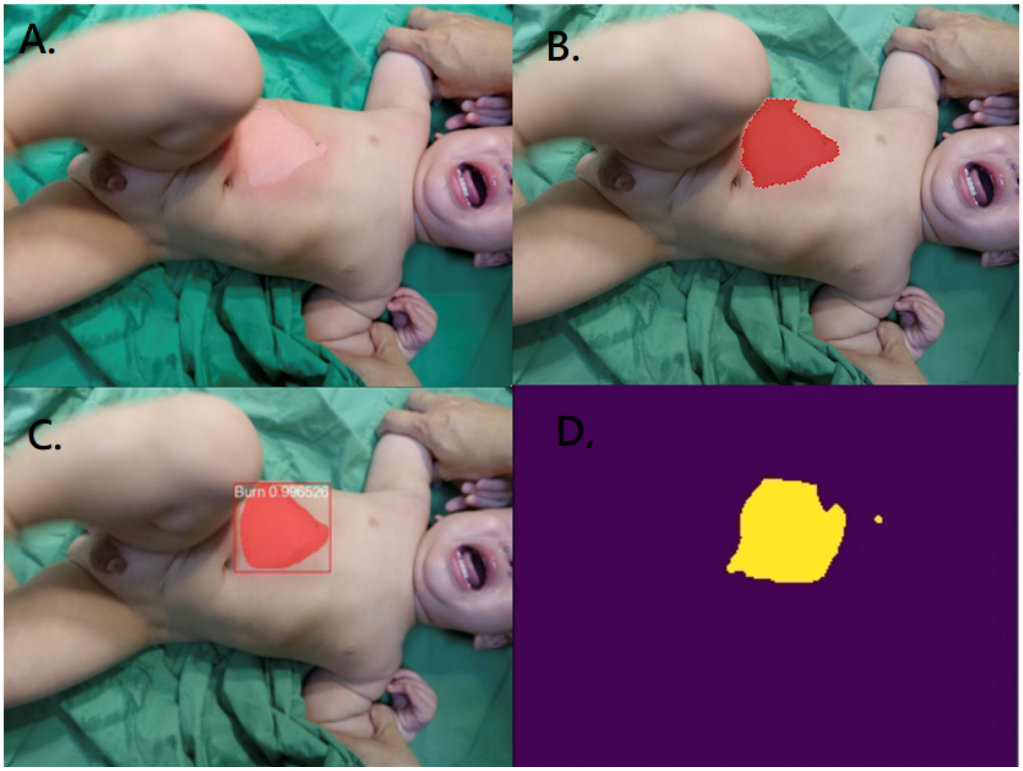


Figure 3. Deep partial burn. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.

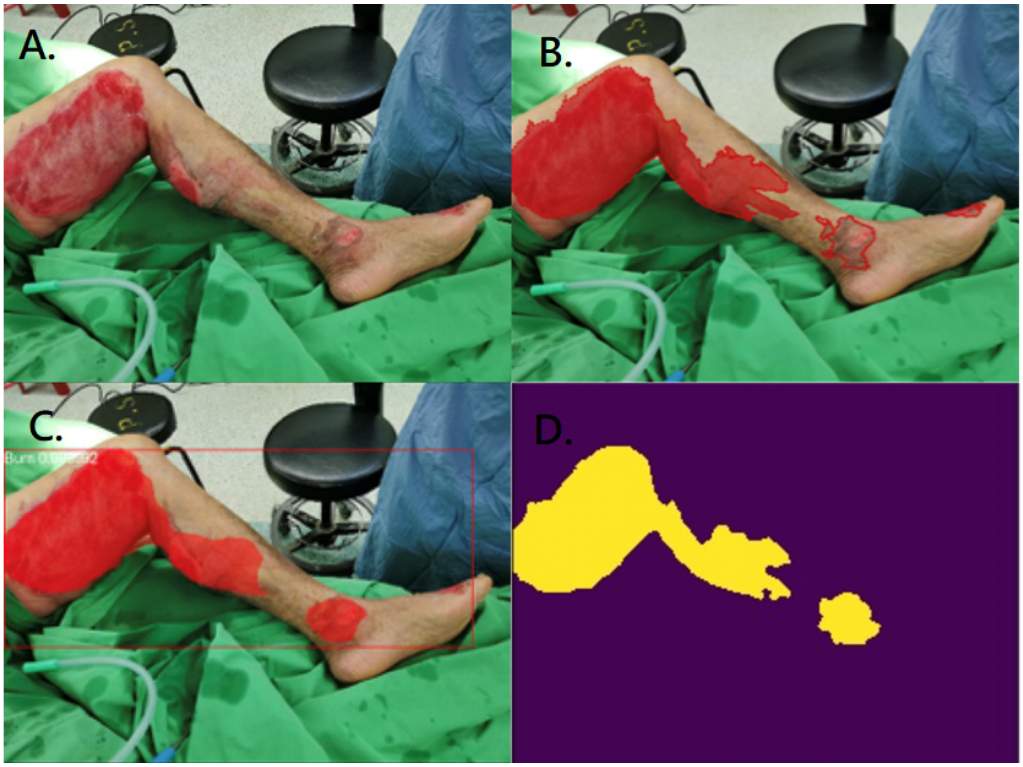


Figure 4. Full thickness burn. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.

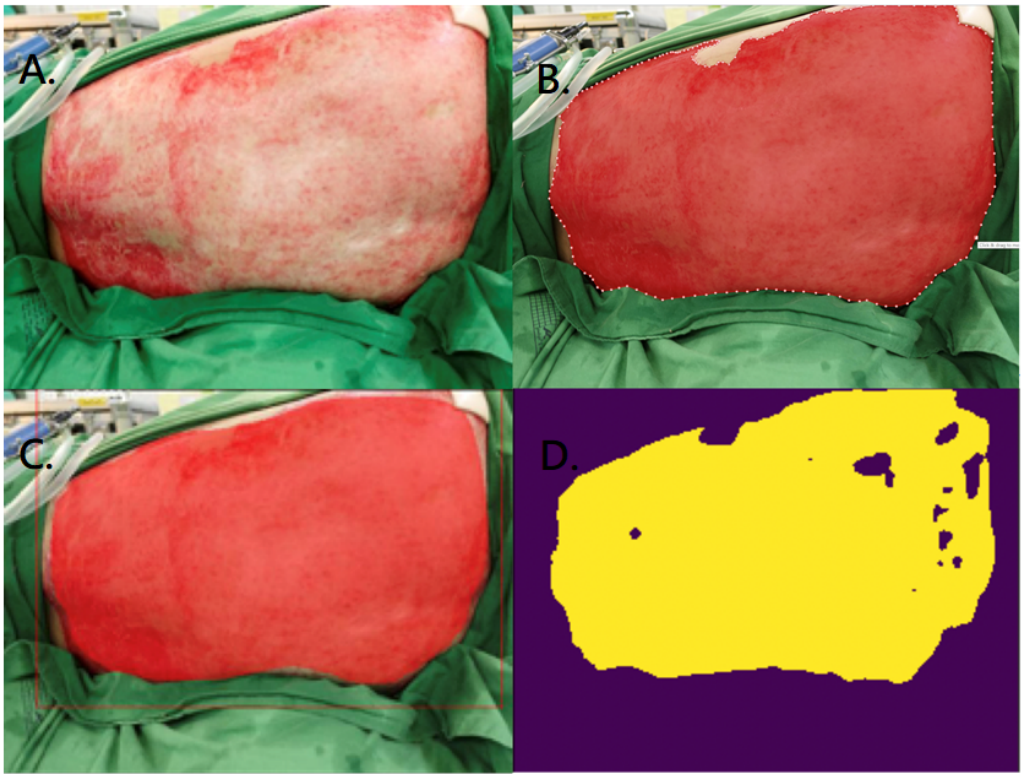
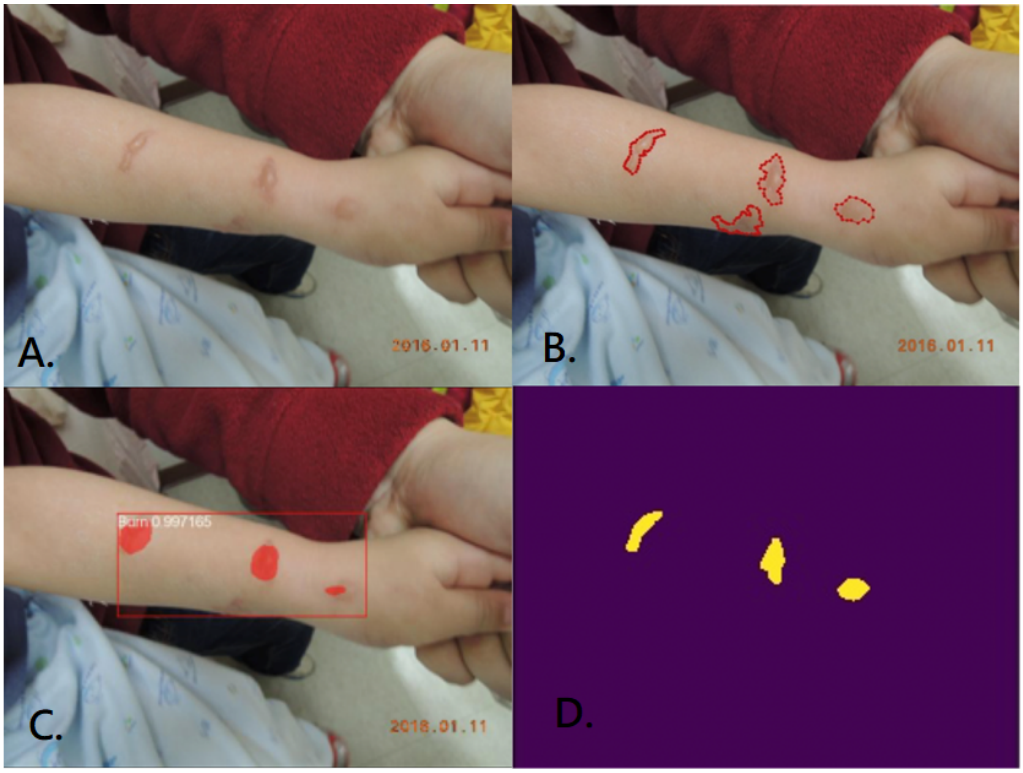


Figure 5. Small scattered burns. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.



Segmentation of Hands and Palms

A total of 400 images of both volar hands were collected and labeled. The male-to-female ratio was 193:207. Since U-Net and Mask R-CNN both performed better with the ResNet101 backbone than with the ResNet50 backbone in the burn wound segmentation, only ResNet101 was applied in the segmentation of the hand and palm data sets.

Contrary to the burn wound results, U-Net had slightly better overall performance in the segmentation of the hands and palms than Mask R-CNN (Table 5 and Table 6). For hand segmentation, U-Net had a DC of 0.9920 and Mask R-CNN had a DC of 0.9692. For palm segmentation, the difference was not as obvious with a DC of 0.9910 versus 0.9803. Figure 6 provides a representative example of the segmentation of a particular hand by both U-Net and Mask R-CNN, while Multimedia Appendix 2 provides an example for a palm.

Table 5. Segmentation results for hands with ResNet101.

Variable	U-Net	Mask R-CNN
Mean DC[a]	0.9920	0.9692
Mean IoU[b]	0.9842	0.9405
Mean precision	0.9906	0.9657
Mean recall	0.9935	0.9728
Mean accuracy	0.9933	0.9407

[a]DC: Dice coefficient.

[b]IoU: intersection over union.

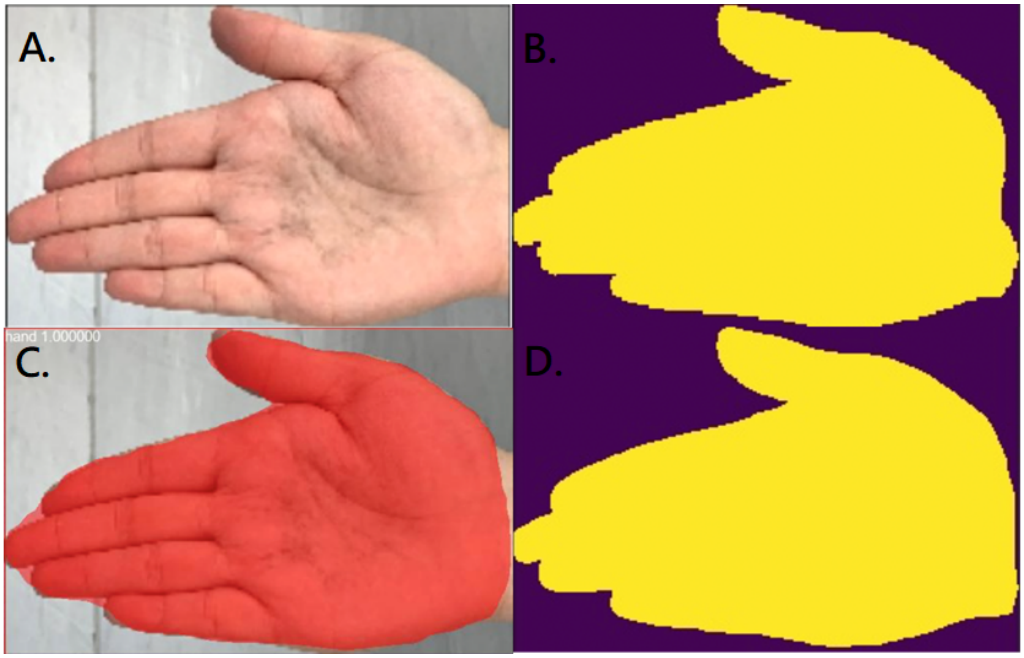
Table 6. Segmentation results for palms with ResNet101.

Variable	U-Net	Mask R-CNN
Mean DC[a]	0.9910	0.9803
Mean IoU[b]	0.9822	0.9614
Mean precision	0.9904	0.9836
Mean recall	0.9916	0.9770
Mean accuracy	0.9878	0.9615

[a]DC: Dice coefficient.

[b]IoU: intersection over union.

Figure 6. Segmentation of the hand. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.



Burn Segmentation to %TBSA

In the last part of our study, we designed a test to compare the estimation of the percentage of TBSA burned according to surgeons and Mask R-CNN. Photos of the abdomen, left thigh, left leg, right leg, and left hand of a patient were taken from the same distance (Figure 7). Images of the burn wounds and of the hands were co-labeled by 2 surgeons as ground truth. The previously trained Mask R-CNN with the ResNet101 backbone was used to calculate the %TBSA of each wound. Then, pictures of the burn wounds and the hands were given to 5 burn surgeons, and they gave their respective estimations of %TBSA. The results of each surgeon, ground truth, and Mask R-CNN are shown in Multimedia Appendix 3. The ground truth was a pixel-based calculation (abdomen: 2.07%, thigh: 2.06%, right leg and knee: 2.64%, and left leg: 2.85%). Mask R-CNN had a smaller average deviation (0.115% TBSA) from ground truth than all of the burn surgeons (0.45%-1.14% TBSA; Figure 8).

Figure 7. A1-A5: original image of the left hand, abdomen, left thigh, right leg, and left leg. B1-B5: labeled images as ground truth.

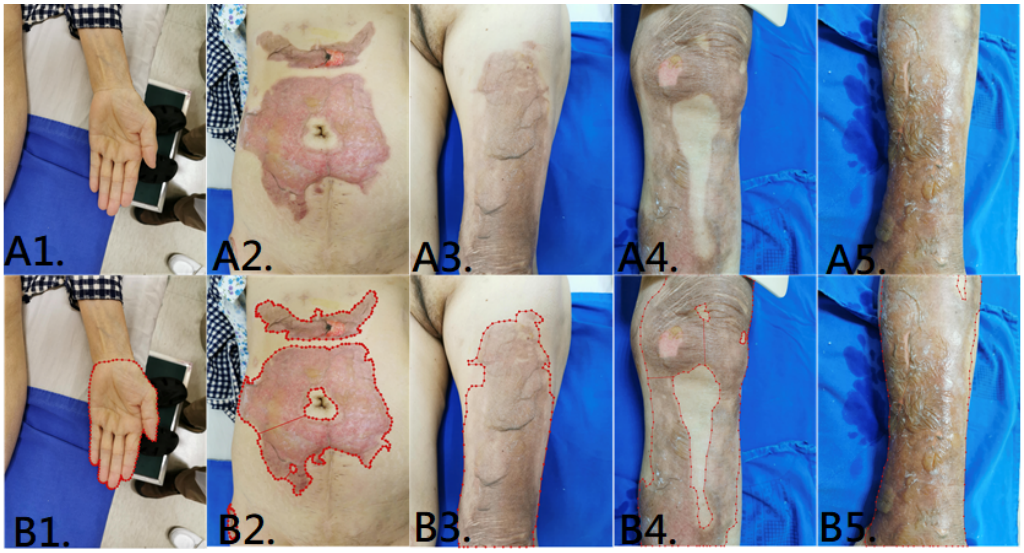
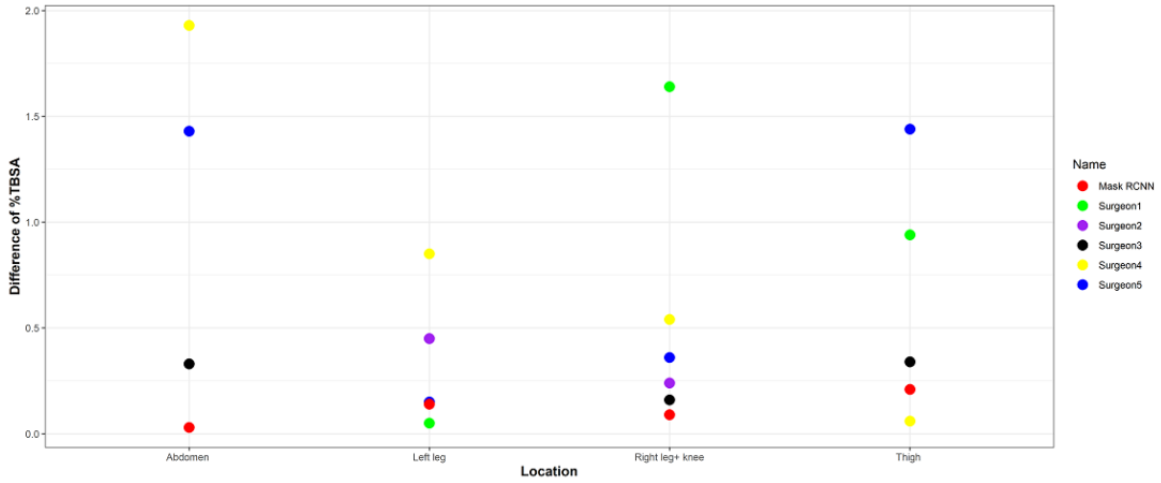


Figure 8. Differences between ground truth and estimated %TBSA of Mask R-CNN and burn surgeons at various burn sites. %TBSA: percentage total body surface area.



Discussion

Data Sets

Studies of machine learning in burn diagnosis are relatively rare, because there are challenges in establishing accurate data sets. To begin with, unlike medical images from X-ray or computed tomography (CT) scans, images of burn wounds are not acquired under a standard protocol. Images of burn wounds are acquired using different equipment under various circumstances, such as illumination conditions, distance to the patient, and the background scene. These factors make it difficult to achieve a uniform standard of labeling and annotation.

Next, the numbers of burn images compared with other open image data sets, such as MNIST (70,000 images) and CIFAR-10 (60,000 images), are limited. In recent studies of burn wound segmentation, Despo et al used 656 images for training [23] and Jiao used 1000 images for training [24]. We used 2332 labeled images from all burn depths for training and 259 images for testing. Images of burn wounds are difficult to collect. Unlike cancer imaging archives, there are no high-quality open data sets of images of burn wounds. This may be because complete deidentification of these images is not possible. Researchers are asked not to publish these images as open data sets due to patient privacy. Researchers from different medical facilities are not permitted to share the images with each other as well. Under these circumstances, federated learning to form a global model may be a feasible method to improve the accuracy of different individual models. The concept of federated learning is to share only the weights and bias of different models without sharing data sets [27,28].

In addition, burn wounds, unlike tumors that are detected on magnetic resonance imaging (MRI) images, are not commonly sampled for biopsy to confirm diagnosis. For any pixels on the images, if no other diagnostic technology is used, the true burn depths are hard to ascertain. The images, even when labeled by burn specialists, are relative ground truth only. A given image may receive many different labels when assessed by many doctors.

Finally, many burn wounds have a mixture of several burn depths. If the object of deep learning is to build a burn depth classifier, most images cannot be included for training. Images

of burn wounds require preprocessing as discussed previously in the methods.

In the early work of our study, we tried to build a burn depth classifier. We divided the images of burn wounds into the following 4 categories based on burn depth: superficial (112 images), superficial partial (201 images), deep partial (165 images), and full thickness (170 images). We imported the data set into IBM Visual Insights (previously PowerAI Vision), a tool that can train models to do the classification task. We did data augmentation to enlarge the data set and improve generalization. Then, we chose pretrained GoogLeNet as our network structure. This model showed decent results, with a mean accuracy of 93% (Multimedia Appendix 4). However, some images in the category "superficial partial" had regions with other burn depths as well. The confusion matrix showed more false negative results in this group than in the others (Multimedia Appendix 5). Hence, the accuracy of the model as a burn depth classifier largely depended on the burn wound images collected.

The abovementioned confounding factors also had an impact in previous studies of machine learning used to segment images of burn wounds. In the study by Despo et al, the margins of burn wounds on images were labeled by a surgeon. Then, every image was annotated to 1 severity of burn depth. Since the burn wound depths were not homogeneous, accuracy and IoU were greater in partial thickness burns [23]. In our study, we also faced the same challenges. Initially, every image was labeled by 2 burn surgeons to obtain 2 labeled images. When the burn wounds had multiple burn depths, the labeled areas of the 2 surgeons had more discrepancy. When we input the discrepantly labeled images to train the models, they resulted in a good mask of the overall burn area but an incorrect classification of burn depth segmentation (Multimedia Appendix 6 and Multimedia Appendix 7). Zhang et al reported an interesting finding [29]. When they input randomly labeled objects or random pixels, after 10,000 steps, their neural network models still converged to fit the training set perfectly. The neural networks were rich enough to memorize bad training data. Yet, their results on testing data sets were poor. To avoid the problem of ambiguous ground truth, we modified the method so that only the burn wound margin was co-labeled by the 2 burn surgeons. This was because the ground truth of the margins had the highest consensus and because all formulae used for burn resuscitation only involved total burn area, which is equivalent to burn margin and is not related to burn depth.

Segmentation Results

We chose U-Net and Mask R-CNN as our main models for segmentation of burn wounds and hands because they are both popular and well-developed CNN models. Although they have different architectures and use different loss functions, their segmentation output seems similar. U-Net outputs semantic segmentation, and it is the most common segmentation model in the medical field [30]. U-Net has been deployed in the evaluation of various sources of medical images, such as positron emission tomography (PET) scans of brain lesions [31], microscopy images of cells [32], CT scans of thoracic organs [33], and MRI scans of breast lesions [34]. Mask R-CNN was developed by Facebook AI Research, and it outputs object detection with instance segmentation [35,36]. Mask R-CNN began getting attention in the medical field in 2018. It has been deployed in the analysis of various sources of medical images as well, such as PET scans of lung lesions [37], sonographic images of breast lesions [38], and MRI scans of knee injuries [39].

Previous studies have also applied these 2 models. Vuola et al reported a study of nuclei segmentation of microscopy images. U-Net had a better DC and created more accurate segmentation masks. Mask R-CNN had better recall and precision, and could detect nuclei more accurately but struggled to predict a good segmentation mask [40]. Zhao et al reported a study of tree canopy segmentation of aerial images. Mask R-CNN performed better in segmentation as well as in tree detection [41]. Bouget et al reported a study of thoracic structure segmentation combining 2 models. Mask R-CNN had the weakness of underestimating structural boundaries, and it required a longer training time. U-Net had the weakness of spatial inconsistency when compiling 2D segmentation results into 3D [42]. In our study, Mask R-CNN was better at burn wound segmentation, while U-Net was better at hand segmentation. We believe that when the segmented objects have similar shape and size, such as with nuclei, hands, and palms, U-Net can achieve better segmentation results than Mask R-CNN. Mask R-CNN had to take into account the loss function components from estimating the bounding box and class, not just the mask. The weights of the bounding box and class components are calculated prior to the weight of the mask component in order to get accurate instance location. Huang et al proposed a modified Mask R-CNN to improve mask prediction [43].

However, the performance of U-Net in burn wound segmentation was not as good as that of Mask R-CNN. The burn wounds comprised 3 types of burn depths with various colors, hues, and textures, and were also of irregular shape and different sizes. Because it lacks the RPN function of Mask R-CNN, U-Net may not have the volume to "memorize" all the features of burn wounds by convolution and de-convolution. In the Kaggle science bowl, both U-Net and Mask R-CNN achieved excellent results after fine tuning. Hence, the performance of the 2 models may depend on the segmentation task, the data sets, and fine tuning.

The segmentation result is not the only consideration. There are other comparative pros and cons of these 2 models. If a model is deployed in mobile devices, time consumption for prediction is an important factor. In our study, it took less time for U-Net (0.035 s/image) to do the prediction than for Mask R-CNN (0.175 s/image). The total time needed to train Mask R-CNN was about 1.5 times that needed to train U-Net. In addition, semantic segmentation involves direct pixel classification. If the objective is to calculate the total burn area, U-Net is capable of producing good results. If we want to segment different types of wounds on the same images, such as incisions and abrasions, Mask R-CNN can provide classification confidence in each of the RoIs, not just the masks.

Both U-Net and Mask R-CNN can segment burn wounds of any burn depths (Figures 2-4). The segmentation result was

more satisfactory when areas were large and confluent (Figure 4). If the burn wound (pixels) was small, the segmentation results of both models were not satisfactory (Figure 5). This is because a small area is susceptible to resizing, convolution, and max pooling. Similar observations were reported by Bouget et al, when they segmented structures inside the chest wall [42]. Large structures, such as the heart, lungs, and spine, had a DC of more than 0.95. Small structures, such as lymph nodes, had a DC of only around 0.41. In the study by Vuola et al, they removed the very small masks (under 10 pixels) to improve the prediction [40]. Fortunately, small and scattered burns are less critical clinically.

Conversion of Segmentation Mask to %TBSA

There exist other methods for converting a segmentation mask to %TBSA. One approach is to acquire the actual burn area (eg, 225 cm^2) by calculating the relation of pixels of the mask area on the image and the distance from the wound to the camera. The next step is to calculate the body surface area (BSA; eg, 17,525 cm^2) via the patient's body weight, height, and gender. The %TBSA of the burn wound can be calculated by dividing these 2 numbers. Although this approach seems straightforward, there are more than 25 formulae to estimate BSA based on studies of different populations [44]. When it comes to child BSA, we need completely different formulae for calculation, again with various degrees of accuracy [45].

We adopted the rule of hand/palm as a guide to estimate %TBSA, because the rule of hand/palm shows very little difference between racial groups, genders, BMI, and ages [8,46]. The rule of hand/palm can also be used in children and infants, where it is closer to the original 1% TBSA rule. Moreover, thumbprints, which are approximately 1/30 TBSA, can also be used as a guide to estimate areas of small burns [47]. In our study, only 17 images were burn injuries involving the volar hand. We therefore collected images of healthy hands from our colleagues rather than using burned hands to train the models.

In the last stage of our study, we conducted a test to compare the %TBSA estimated by burn surgeons and by Mask R-CNN with a ResNet101 backbone. Mask R-CNN had less variance from ground truth on average. It is very important to have a small deviation on every estimation. If a patient has multiple burn sites, the errors from each wound may add up to become a large deviation. In a study by Parvizi et al, the difference in estimation by inspection across burn experts was found to be as large as 16.5% TBSA in an adult patient and 31.5% TBSA in a child patient, which resulted in great volume differences in the estimation of fluid needed for resuscitation [10]. Our method was aimed to derive similar estimates when the same burn wound was estimated by different burn experts by inspection, such as by teleconsultation. In reality, burn surgeons would typically visit patients and calculate the area more meticulously. Additionally, the burn area would be recalculated in the days following the burn injury. Theoretically, the variability among estimations would be less than when the burn area is estimated just by inspecting an image of the burn wound.

Limitations

The data set of burn wounds was collected from a single medical center in Taiwan. Although it is currently the largest data set, the number of training images was small. The models require more input images to improve accuracy.

Our deep learning models can segment a burn wound of any burn depth. However, they are unable to classify burn depths on segmentation. This is so because the ground truth of burn depths is hard to define by burn surgeons consistently. Further study may apply machine learning to assist in burn depth labeling before input for training.

We used normal hands as a template to calculate the %TBSA burned. When a patient had burns involving both hands, our models could still segment the burned hands. Since children's hands are shaped similarly to those of adults, our models can presumably also segment the hands of children (Multimedia Appendix 8). However, we did not collect enough images to directly assess accuracy in these circumstances.

Our data set did not include burn wounds from patients with markedly different skin tones. We hypothesize that the deep learning models will accurately detect burn wounds when the burn injury is more severe than superficial second degree, where the skin layers that are deeper than the pigment cells are disrupted. For example, a superficial second-degree burn injury with ruptured bullae shows a similar shade of pink even on different skin tones. Yet, skin tone will definitely contribute to the performance of the models. Convolution layers and the RoI obtained by deep learning largely depend on the relationship with their adjacent pixels. To test our hypothesis, we collected 100 web scraping images of burn wounds from different skin tones and input them into our models for wound segmentation (Multimedia Appendix 9). The results confirmed that our models performed well when the burn injury was more severe than superficial second degree. However, the segmentation results varied when the burn wound had no bullae formation or rupture (whether superficial second or first degree). To resolve this problem, we need more quality images to correlate skin tone with segmentation performance.

Finally, burn wound images are 2D projections of 3D burn wounds, akin to the Mercator world map. Unlike the world map, the cross sections of the trunk and extremities of the human body are not just ellipses or circles. The distance of the camera from the wound bed can be adjusted for by a simple formula, but adjusting for the angle at which the photos are taken requires complex differential and integral formulae with multiple variables. To get the most accurate estimation of %TBSA, we suggest taking all photos at a constant distance of around 30 to 50 cm and holding the camera (cellphone) parallel to the wound bed to decrease the effect of the angle. Our study will further deploy models on images taken with a 3D camera to acquire more accurate results.

Conclusions

To the best of our knowledge, this is the first study to determine the %TBSA of burn wounds with different deep learning models. Based on the rule of hand, %TBSA can be calculated by comparing segmentation masks of the burn wound and hand

of a patient. In our study, Mask R-CNN with ResNet101 performed this task satisfactorily in comparison with burn surgeons. With the assistance of deep learning, the fluid resuscitation and nutritional needs of burn injury patients can be more precisely and accurately assessed.

Acknowledgments

This work was supported by the Innovation Project of Far Eastern Memorial Hospital (grant number PI20200002). We thank our colleagues in the Department of Surgery and the Department of Nursing (operating room, 13G ward) of Far Eastern Memorial Hospital and in the Graduate Institute of Biomedical Electronics & Bioinformatics of National Taiwan University for the collection of the images of hands. We also thank Shih-Chen Huang, who helped coordinate with the burn surgeons in the collection of the label data.

Conflicts of Interest

None declared.

Multimedia Appendix 1
U-Net architecture, encoder, and decoder replaced with ResNet.
[PNG File , 107 KB - medinform_v9i12e22798_app1.png]

Multimedia Appendix 2
Segmentation of the palm. A: original photo; B: ground truth; C: result of Mask R-CNN; D: result of U-Net.
[PNG File , 1665 KB - medinform_v9i12e22798_app2.png]

Multimedia Appendix 3
Estimation of %TBSA burned according to 5 different burn surgeons (ground truth and Mask R-CNN). %TBSA: percentage total body surface area.
[PNG File , 150 KB - medinform_v9i12e22798_app3.png]

Multimedia Appendix 4
Mean accuracy of burn depth classification.
[PNG File , 114 KB - medinform_v9i12e22798_app4.png]

Multimedia Appendix 5
Confusion matrix of different subgroups.
[PNG File , 99 KB - medinform_v9i12e22798_app5.png]

Multimedia Appendix 6
A mix of all burn depths in a burn wound.
[PNG File , 1407 KB - medinform_v9i12e22798_app6.png]

Multimedia Appendix 7
Incorrect prediction of burn depths but correct prediction of total burn area.
[PNG File , 260 KB - medinform_v9i12e22798_app7.png]

Multimedia Appendix 8
A: adult hand burn; B: segmentation of adult hand burn; C: child hand burn; D: segmentation of child hand burn.
[PNG File , 1732 KB - medinform_v9i12e22798_app8.png]

Multimedia Appendix 9
Web scraping images of burn wounds on patients with markedly lighter and darker skin tone in comparison to our study population.
[PNG File , 1338 KB - medinform_v9i12e22798_app9.png]

References

1. Harish V, Raymond AP, Issler AC, Lajevardi SS, Chang L, Maitz PK, et al. Accuracy of burn size estimation in patients transferred to adult Burn Units in Sydney, Australia: an audit of 698 patients. Burns 2015 Feb;41(1):91-99. [doi: 10.1016/j.burns.2014.05.005] [Medline: 24972983]
2. Baartmans M, van Baar M, Boxma H, Dokter J, Tibboel D, Nieuwenhuis M. Accuracy of burn size assessment prior to arrival in Dutch burn centres and its consequences in children: a nationwide evaluation. Injury 2012 Sep;43(9):1451-1456. [doi: 10.1016/j.injury.2011.06.027] [Medline: 21741042]
3. Resch TR, Drake RM, Helmer SD, Jost GD, Osland JS. Estimation of burn depth at burn centers in the United States. Journal of Burn Care & Research 2014;35(6):491-497. [doi: 10.1097/bcr.0000000000000031]
4. Jaskille AD, Shupp JW, Jordan MH, Jeng JC. Critical review of burn depth assessment techniques: Part I. Historical review. Journal of Burn Care & Research 2009;30(6):937-947. [doi: 10.1097/bcr.0b013e3181c07f21]
5. Monstrey S, Hoeksema H, Verbelen J, Pirayesh A, Blondeel P. Assessment of burn depth and burn wound healing potential. Burns 2008 Sep;34(6):761-769. [doi: 10.1016/j.burns.2008.01.009] [Medline: 18511202]
6. Jaspers ME, van Haasterecht L, van Zuijlen PP, Mokkink LB. A systematic review on the quality of measurement techniques for the assessment of burn wound depth or healing potential. Burns 2019 Mar;45(2):261-281. [doi: 10.1016/j.burns.2018.05.015] [Medline: 29941159]
7. Thatcher JE, Squiers JJ, Kanick SC, King DR, Lu Y, Wang Y, et al. Imaging techniques for clinical burn assessment with a focus on multispectral imaging. Adv Wound Care (New Rochelle) 2016 Aug 01;5(8):360-378 [FREE Full text] [doi: 10.1089/wound.2015.0684] [Medline: 27602255]
8. Thom D. Appraising current methods for preclinical calculation of burn size - A pre-hospital perspective. Burns 2017 Feb;43(1):127-136. [doi: 10.1016/j.burns.2016.07.003] [Medline: 27575669]
9. Neaman KC, Andres LA, McClure AM, Burton ME, Kemmeter PR, Ford RD. A new method for estimation of involved BSAs for obese and normal-weight patients with burn injury. Journal of Burn Care & Research 2011;32(3):421-428. [doi: 10.1097/bcr.0b013e318217f8c6]
10. Parvizi D, Kamolz L, Giretzlehner M, Haller HL, Trop M, Selig H, et al. The potential impact of wrong TBSA estimations on fluid resuscitation in patients suffering from burns: things to keep in mind. Burns 2014 Mar;40(2):241-245. [doi: 10.1016/j.burns.2013.06.019] [Medline: 24050977]
11. Kwon S, Hong J, Choi E, Lee B, Baik C, Lee E, et al. Detection of atrial fibrillation using a ring-type wearable device (CardioTracker) and deep learning analysis of photoplethysmography signals: prospective observational proof-of-concept study. J Med Internet Res 2020 May 21;22(5):e16443 [FREE Full text] [doi: 10.2196/16443] [Medline: 32348254]
12. Adam G, Rampášek L, Safikhani Z, Smirnov P, Haibe-Kains B, Goldenberg A. Machine learning approaches to drug response prediction: challenges and recent progress. NPJ Precis Oncol 2020 Jun 15;4(1):19 [FREE Full text] [doi: 10.1038/s41698-020-0122-1] [Medline: 32566759]
13. Kanavati F, Toyokawa G, Momosaki S, Rambeau M, Kozuma Y, Shoji F, et al. Weakly-supervised learning for lung carcinoma classification using deep learning. Sci Rep 2020 Jun 09;10(1):9297 [FREE Full text] [doi: 10.1038/s41598-020-66333-x] [Medline: 32518413]
14. Han W, Johnson C, Gaed M, Gómez JA, Moussa M, Chin JL, et al. Histologic tissue components provide major cues for machine learning-based prostate cancer detection and grading on prostatectomy specimens. Sci Rep 2020 Jun 18;10(1):9911 [FREE Full text] [doi: 10.1038/s41598-020-66849-2] [Medline: 32555410]
15. Kanevsky J, Corban J, Gaster R, Kanevsky A, Lin S, Gilardino M. Big data and machine learning in plastic surgery. Plastic and Reconstructive Surgery 2016;137(5):890e-897e. [doi: 10.1097/prs.0000000000002088]
16. Liu NT, Salinas J. Machine learning in burn care and research: A systematic review of the literature. Burns 2015 Dec;41(8):1636-1641. [doi: 10.1016/j.burns.2015.07.001] [Medline: 26233900]
17. Serrano C, Acha B, Gómez-Cía T, Acha JI, Roa LM. A computer assisted diagnosis tool for the classification of burns by depth of injury. Burns 2005 May;31(3):275-281. [doi: 10.1016/j.burns.2004.11.019] [Medline: 15774281]
18. Acha B, Serrano C, Acha JI, Roa LM. Segmentation and classification of burn images by color and texture information. J Biomed Opt 2005;10(3):034014 [FREE Full text] [doi: 10.1117/1.1921227] [Medline: 16229658]
19. Acha B, Sonka M, Serrano C, Palencia S, Murillo J. Classification of burn wounds using support vector machines. 2004 Presented at: Medical Imaging 2004; May 12, 2004; San Diego, CA. [doi: 10.1117/12.535491]
20. Acha B, Serrano C, Fondon I, Gomez-Cia T. Burn depth analysis using multidimensional scaling applied to psychophysical experiment data. IEEE Trans. Med. Imaging 2013 Jun;32(6):1111-1120. [doi: 10.1109/tmi.2013.2254719]
21. Serrano C, Boloix-Tortosa R, Gómez-Cía T, Acha B. Features identification for automatic burn classification. Burns 2015 Dec;41(8):1883-1890. [doi: 10.1016/j.burns.2015.05.011] [Medline: 26188898]
22. Cirillo M, Mirdell R, Sjöberg F, Pham T. Time-independent prediction of burn depth using deep convolutional neural networks. J Burn Care Res 2019 Oct 16;40(6):857-863. [doi: 10.1093/jbcr/irz103] [Medline: 31187119]
23. Despo O, Yeung S, Jopling J, Pridgen B, Sheckter C, Silberstein S, et al. BURNED: Towards Efficient and Accurate Burn Prognosis Using Deep Learning. 2017. URL: http://cs231n.stanford.edu/reports/2017/pdfs/507.pdf [accessed 2020-08-17]
24. Jiao C, Su K, Xie W, Ye Z. Burn image segmentation based on Mask Regions with convolutional neural network deep learning framework: more accurate and more convenient. Burns Trauma 2019;7:6 [FREE Full text] [doi: 10.1186/s41038-018-0137-9] [Medline: 30859107]

25. Giretzlehner M, Dirnberger J, Owen R, Haller H, Lumenta D, Kamolz L. The determination of total burn surface area: How much difference? Burns 2013 Sep;39(6):1107-1113. [doi: 10.1016/j.burns.2013.01.021] [Medline: 23566430]
26. Parvizi D, Giretzlehner M, Wurzer P, Klein LD, Shoham Y, Bohanon FJ, et al. BurnCase 3D software validation study: Burn size measurement accuracy and inter-rater reliability. Burns 2016 Mar;42(2):329-335. [doi: 10.1016/j.burns.2016.01.008] [Medline: 26839051]
27. Thapa C, Chamikara M, Camtepe S. Advancements of federated learning towards privacy preservation: from federated learning to split learning. In: Rehman MH, Gaber MM, editors. Federated Learning Systems. Studies in Computational Intelligence, vol 965. Cham: Springer; 2021:79-109.
28. Yang D, Xu Z, Li W, Myronenko A, Roth HR, Harmon S, et al. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med Image Anal 2021 May;70:101992 [FREE Full text] [doi: 10.1016/j.media.2021.101992] [Medline: 33601166]
29. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. Commun. ACM 2021 Mar;64(3):107-115. [doi: 10.1145/3446776]
30. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Cham: Springer; 2015:234-241.
31. Blanc-Durand P, Van Der Gucht A, Schaefer N, Itti E, Prior JO. Automatic lesion detection and segmentation of 18F-FET PET in gliomas: A full 3D U-Net convolutional neural network study. PLoS One 2018 Apr 13;13(4):e0195798 [FREE Full text] [doi: 10.1371/journal.pone.0195798] [Medline: 29652908]
32. Fabijańska A. Segmentation of corneal endothelium images using a U-Net-based convolutional neural network. Artif Intell Med 2018 Jun;88:1-13. [doi: 10.1016/j.artmed.2018.04.004] [Medline: 29680687]
33. Dong X, Lei Y, Wang T, Thomas M, Tang L, Curran WJ, et al. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. Med Phys 2019 May 22;46(5):2157-2168 [FREE Full text] [doi: 10.1002/mp.13458] [Medline: 30810231]
34. Zhang Y, Chen J, Chang K, Park VY, Kim MJ, Chan S, et al. Automatic breast and fibroglandular tissue segmentation in breast MRI using deep learning by a fully-Convolutional Residual Neural Network U-Net. Acad Radiol 2019 Nov;26(11):1526-1535 [FREE Full text] [doi: 10.1016/j.acra.2019.01.012] [Medline: 30713130]
35. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017 Jun;39(6):1137-1149. [doi: 10.1109/TPAMI.2016.2577031] [Medline: 27295650]
36. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice, Italy p. 2980-2988. [doi: 10.1109/iccv.2017.322]
37. Zhang R, Cheng C, Zhao X, Li X. Multiscale Mask R-CNN-Based lung tumor detection using PET imaging. Mol Imaging 2019 Jul 31;18:1536012119863531 [FREE Full text] [doi: 10.1177/1536012119863531] [Medline: 31364467]
38. Chiao J, Chen K, Liao KY, Hsieh P, Zhang G, Huang T. Detection and classification the breast tumors using mask R-CNN on sonograms. Medicine 2019;98(19):e15200. [doi: 10.1097/md.0000000000015200]
39. Couteaux V, Si-Mohamed S, Nempont O, Lefevre T, Popoff A, Pizaine G, et al. Automatic knee meniscus tear detection and orientation classification with Mask-RCNN. Diagn Interv Imaging 2019 Apr;100(4):235-242 [FREE Full text] [doi: 10.1016/j.diii.2019.03.002] [Medline: 30910620]
40. Vuola A, Akram S, Kannala J. Mask-RCNN and U-Net Ensembled for Nuclei Segmentation. 2019 Presented at: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); April 8-11, 2019; Venice, Italy p. 208-212. [doi: 10.1109/isbi.2019.8759574]
41. Zhao T, Yang Y, Niu H, Wang D, Chen Y. Comparing U-Net convolutional network with mask R-CNN in the performances of pomegranate tree canopy segmentation. 2018 Presented at: Proc. SPIE 10780, Multispectral, Hyperspectral, and Ultraspectral Remote Sensing Technology, Techniques and Applications VII; December 21, 2018; Honolulu, HI. [doi: 10.1117/12.2325570]
42. Bouget D, Jørgensen A, Kiss G, Leira HO, Langø T. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. Int J Comput Assist Radiol Surg 2019 Jun;14(6):977-986. [doi: 10.1007/s11548-019-01948-8] [Medline: 30891655]
43. Huang Z, Huang L, Gong Y, Huang C, Wang X. Mask Scoring R-CNN. 2019 Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); June 15-20, 2019; Long Beach, CA p. 6402-6411. [doi: 10.1109/CVPR.2019.00657]
44. Redlarski G, Palkowski A, Krawczuk M. Body surface area formulae: an alarming ambiguity. Sci Rep 2016 Jun 21;6(1):27966 [FREE Full text] [doi: 10.1038/srep27966] [Medline: 27323883]
45. Rumpf RW, Stewart WC, Martinez SK, Gerrard CY, Adolphi NL, Thakkar R, et al. Comparison of the Lund and Browder table to computed tomography scan three-dimensional surface area measurement for a pediatric cohort. J Surg Res 2018 Jan;221:275-284 [FREE Full text] [doi: 10.1016/j.jss.2017.08.019] [Medline: 29229139]
46. Cox S, Kriho K, De Klerk S, van Dijk M, Rode H. Total body and hand surface area: Measurements, calculations, and comparisons in ethnically diverse children in South Africa. Burns 2017 Nov;43(7):1567-1574. [doi: 10.1016/j.burns.2017.04.012] [Medline: 28473269]

47. Dargan D, Mandal A, Shokrollahi K. Hand burns surface area: A rule of thumb. Burns 2018 Aug;44(5):1346-1351. [doi: 10.1016/j.burns.2018.02.011] [Medline: 29534883]

Abbreviations

%TBSA: percentage total body surface area
BSA: body surface area
CNN: convolutional neural network
CT: computed tomography
DC: Dice coefficient
IoU: intersection over union
LDI: laser Doppler imaging
MRI: magnetic resonance imaging
PET: positron emission tomography
RoI: region of interest
RPN: regional proposal network
SVM: support vector machine

Edited by R Kukafka, G Eysenbach; submitted 12.08.20; peer-reviewed by K Ahmad, S Shams; comments to author 06.12.20; revised version received 19.12.20; accepted 15.10.21; published 02.12.21.

Please cite as:
Chang CW, Lai F, Christian M, Chen YC, Hsu C, Chen YS, Chang DH, Roan TL, Yu YC
Deep Learning–Assisted Burn Wound Diagnosis: Diagnostic Model Development Study
JMIR Med Inform 2021;9(12):e22798
URL: https://medinform.jmir.org/2021/12/e22798
doi:10.2196/22798
PMID:34860674

Original Paper

A BERT-Based Generation Model to Transform Medical Texts to SQL Queries for Electronic Medical Records: Model Development and Validation

Youcheng Pan[1], MEng; Chenghao Wang[1], MSc; Baotian Hu[1], PhD; Yang Xiang[2], PhD; Xiaolong Wang[1], PhD; Qingcai Chen[1,2], PhD; Junjie Chen[1], PhD; Jingcheng Du[3], PhD

[1]Intelligent Computing Research Center, Harbin Institute of Technology, Shenzhen, China

[2]Peng Cheng Laboratory, Shenzhen, China

[3]University of Texas Health Science Center at Houston, Houston, TX, United States

Corresponding Author:
Baotian Hu, PhD
Intelligent Computing Research Center
Harbin Institute of Technology
No. 6, Pingshan 1st Road
Shenzhen, 518055
China
Phone: 86 136 9164 0856
Email: hubaotian@hit.edu.cn

Abstract

Background: Electronic medical records (EMRs) are usually stored in relational databases that require SQL queries to retrieve information of interest. Effectively completing such queries can be a challenging task for medical experts due to the barriers in expertise. Existing text-to-SQL generation studies have not been fully embraced in the medical domain.

Objective: The objective of this study was to propose a neural generation model that can jointly consider the characteristics of medical text and the SQL structure to automatically transform medical texts to SQL queries for EMRs.

Methods: We proposed a medical text–to-SQL model (MedTS), which employed a pretrained Bidirectional Encoder Representations From Transformers model as the encoder and leveraged a grammar-based long short-term memory network as the decoder to predict the intermediate representation that can easily be transformed into the final SQL query. We adopted the syntax tree as the intermediate representation rather than directly regarding the SQL query as an ordinary word sequence, which is more in line with the tree-structure nature of SQL and can also effectively reduce the search space during generation. Experiments were conducted on the MIMICSQL dataset, and 5 competitor methods were compared.

Results: Experimental results demonstrated that MedTS achieved the accuracy of 0.784 and 0.899 on the test set in terms of logic form and execution, respectively, which significantly outperformed the existing state-of-the-art methods. Further analyses proved that the performance on each component of the generated SQL was relatively balanced and offered substantial improvements.

Conclusions: The proposed MedTS was effective and robust for improving the performance of medical text–to-SQL generation, indicating strong potential to be applied in the real medical scenario.

(JMIR Med Inform 2021;9(12):e32698) doi:10.2196/32698

KEYWORDS

electronic medical record; text-to-SQL generation; BERT; grammar-based decoding; tree-structured intermediate representation

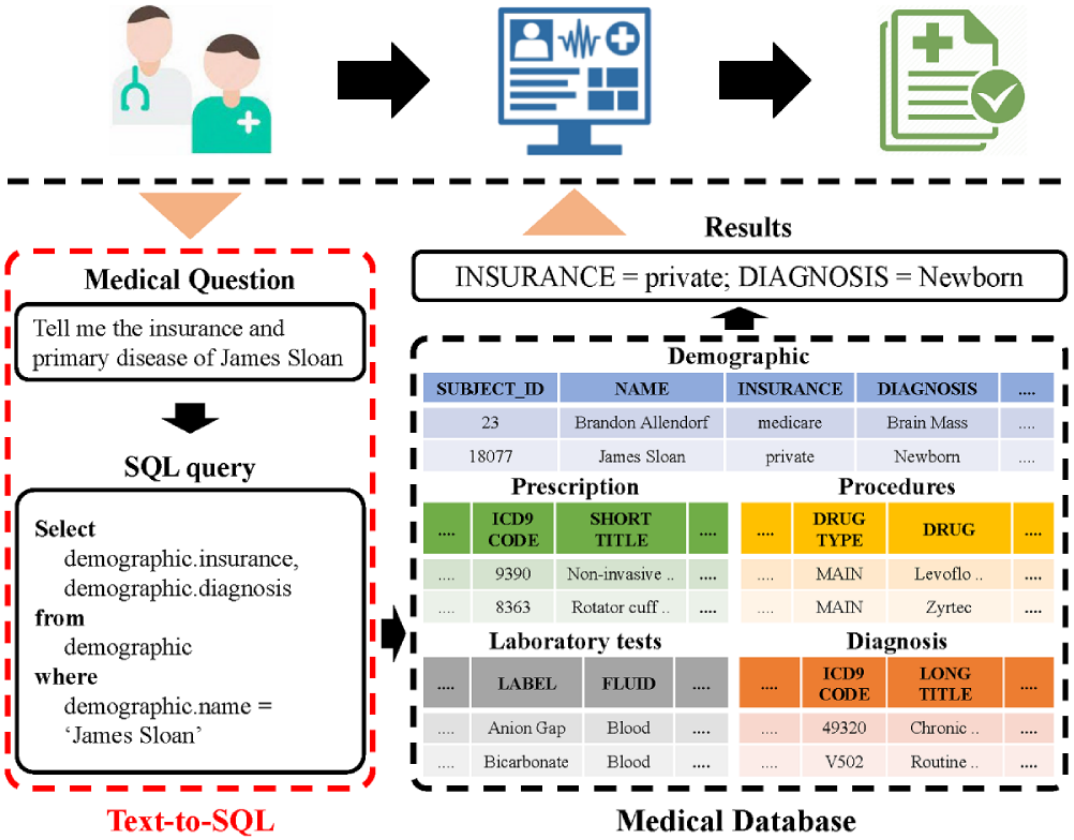
Introduction

Electronic medical records (EMRs) contain abundant medical information on patients and are usually stored in structured relational databases with multiple relational tables [1]. Using EMRs, patient data can be traced back over an extended period of time and by multiple health care providers. EMRs can help identify those who are due for preventive checkups, screenings, or vaccinations. They also can record whether a patient's vital signs (eg, blood pressure, weight) fall within normal limits [2,3].

However, retrieving EMRs from databases may not be easy for medical experts. They usually lack specific training on using SQL to perform queries on relational databases. Even for experienced informaticians, it could be troublesome to deal with massive SQL queries from databases of different structures and applicable scenarios, especially if complex SQL grammars were involved. Therefore, automating the transformation of textual questions written in natural language into SQL queries has great potential to facilitate clinical information retrieval and improve the efficiency of medical diagnosis and treatment decisions.

Text-to-SQL generation [4,5] is the task of transforming natural language questions into SQL queries. As shown in Figure 1, given the medical textual question "Tell me the insurance and primary disease of James Sloan," a text-to-SQL model can transform the question into a SQL query. It is then used to retrieve the corresponding EMR information that is stored in structured medical databases. This task has attracted widespread attention from different domains. The representative studies include automatic terminal information service [6-8] for a flight booking system, GeoQuery [9,10] for a US geography query, WikiSQL [5] for querying Wikipedia, and Spider [11] for realistic applications of several different domains. In many studies, the text-to-SQL generation is regarded as a task similar to natural language generation. Deep neural networks are often adopted as encoders and decoders (eg, the sequence-to-sequence [Seq2Seq] [12] framework with an attention [13,14] or copy mechanism) [15]. The input of the model is the textual question and the output is the SQL query that is viewed as an ordinary word sequence [16,17]. However, the same SQL query can be represented by multiple word sequences, which may affect the training effectiveness of Seq2Seq models. For example, the order of the 2 column names in the Select clause shown in Figure 1 may not influence the execution result of the query, but the Seq2Seq models may treat them as 2 different sequences. To solve this problem, several methods were proposed by incorporating the syntactical structure of SQL [18,19]. For instance, SQLNet [18] proposed a sketch-based sequence-to-set method. A generic sketch highly in line with the SQL grammar was first used and then it only needed to predict the slots in the sketch instead of generating the entire sequence in order.

Figure 1. Application scenario of medical text–to-SQL generation.



Compared with other domains, corresponding explorations in the medical domain are insufficient. Due to the privacy requirements of medical data, a large-scale training corpus is still lacking. Furthermore, jargon and specialized phrases often occur in the medical text. They cannot be represented well by the models trained on other domains. But these terms are sometimes the key points of a medical question. In the limited relevant research, rule-based or those verified on small-scale datasets are most often found, such as methods of translating the medical questions into SPARQL Protocol and RDF Query Language (SPARQL) queries [20] and converting the clinical questions into EMR-dependent structured queries [21]. To push this forward, Yu et al [22] introduced a new criteria-to-SQL generation dataset for clinical trials. However, the targeted free text is quite different from other query text in terms of length and content. Wang et al [23] constructed the first large-scale text-to-SQL generation dataset, MIMICSQL, in the medical domain based on the widely used Medical Information Mart for Intensive Care (MIMIC III) dataset [24]. They also proposed a Seq2Seq-based model, Translate-Edit Model for

Question-to-SQL (TREQS), to directly generate the SQL query for a given medical question by using the dynamic and temporal attention mechanism and controlled copying technique. But these works are preliminary explorations and do not integrate much intrinsic information related to the SQL itself (eg, the tree structure of SQL). Therefore, there is still much room left for further progress.

In this study, we propose a novel model for medical text–to-SQL generation named MedTS for the medical text–to-SQL generation task. First, the medical entities (ie, the table and column names) are recognized via schema linking. A pretrained Bidirectional Encoder Representations From Transformers (BERT) [25] model is then used as an encoder to enhance the question representation. The BERT-based encoder can exploit the relationship of entities between medical text question and database schema. Second, a grammar-based long short-term memory (LSTM) [26] decoder is adopted to generate the tree-structured intermediate representation instead of directly transforming a medical question into SQL query. It is in accordance with the chronological order of the syntax tree of SQL and can reduce the search space at each decoding step. Finally, according to the predefined set of context-free grammar, the intermediate representation is transformed into the corresponding SQL query. Experiments were conducted on the MIMICSQL dataset. We compared the proposed model with 5 competitor methods and further analyzed the performance of each component of the generated SQL query. An online system is accessible to better demonstrate the application of MedTS [27].

Methods

Dataset

We evaluated our proposed method on MIMICSQL [23], which is the first large-scale medical dataset for text-to-SQL generation task in the health care domain. The medical information in MIMICSQL is derived from MIMIC III. All of the medical information was first anonymized to protect patient privacy and then stored in 5 tables in the medical database (Figure 1), including demographic (Demo), laboratory tests (Lab), diagnosis (Diag), procedures (Pro), and prescriptions (Pres). The questions and corresponding SQL queries in MIMICSQL were automatically generated based on fixed templates [28]. Next, 8 freelancers with medical domain knowledge were recruited from a crowd-sourcing platform to validate the question as realistic and reasonable or rephrase the generated question. The information of the MIMICSQL dataset is summarized in Table 1. We adopted the same data partition as in the TREQS [23], in which all the question-SQL pairs were randomly split into training, validation, and test sets in the ratio of 0.8:0.1:0.1, respectively.

Table 1. The summary of the MIMICSQL dataset.

Type	Count
Patients, n	46,520
Tables, n	5
Columns in tables[a], n	23/5/5/7/9
Question-SQL pairs, n	10,000
Template question length (in words), mean	18.39
Rephrased question length (in words), mean	16.45
SQL query length, mean	21.14
Aggregation columns, mean	1.1
Conditions, mean	1.76

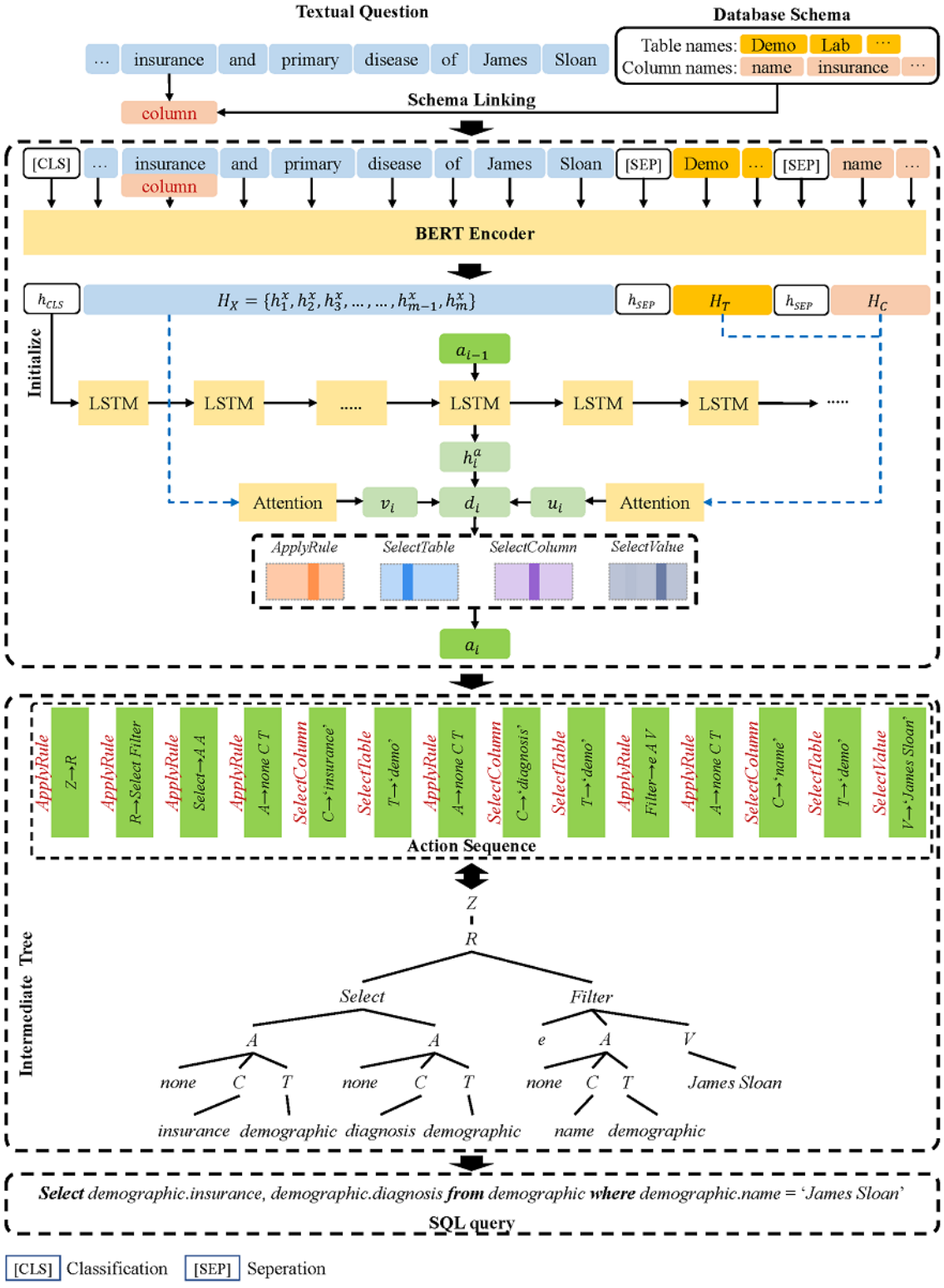
[a]The 23/5/5/7/9 correspond to the numbers of columns in the Demographics/Diagnosis/Procedure/Prescriptions/Laboratory tests tables.

Overview of the Proposed Method

Given a textual question $X=\{x_1,x_2,x_3,...\}$ and the database schema, the goal of this work was to transform the textual question into a SQL query, while ensuring the SQL query retained the same semantic meaning as the textual question.

An overview of MedTS is shown in Figure 2. In the first step, schema linking recognized the database schema information and added corresponding linking marks into the question. Second, the textual question along with linking marks and the database schema information were fed into the pretrained encoder and grammar-based decoder to generate an intermediate representation. Third, the final SQL query was generated based on the intermediate representation.

Figure 2. Overview of our proposed method of medical text–to-SQL task. LSTM: long short-term memory.



Schema Linking

Similar to the method in IRNet [19], the purpose of schema linking was to recognize the mentioned entities in the medical question and assign a linking mark, which referred to recognizing the column names and table names in the medical database. We enumerated all the *n*-grams ($n \in [1,5]$) in a question and arranged them in descending order based on the length. If an *n*-gram exactly matched a column name or was a subset of a column name, we marked this *n*-gram as a column. The recognition of a table followed the same way. If an *n*-gram was recognized as both a column and a table, we marked it as a column because the column mark has higher priority than the table mark. Once an *n*-gram was identified, we removed other n-grams that overlapped with it. By doing this, we obtained all the entities mentioned in the question. Once an entity was recognized and linked with a mark, it became a span and was

encoded into one vector in the encoding process, such as the insurance recognized as a column in Figure 2.

Attention-Based Encoder Using Pretrained BERT

After schema linking, we identified the entities and assigned linking marks. The given medical question X was transformed to $[(x_1,\tau_1),...,(x_m,\tau_m)]$ where x_i was the i^{th} span and τ_i was the mark of x_i assigned during schema linking. If x_i was not an entity, τ_i was *None*. Let $C=\{c_1,c_2,...\}$ and $T=\{t_1,t_2,...\}$ denote the set of all column names and table names. In order to enhance the relationship between the question and database schema, we concatenated the question X and database schema $[C,T]$ with special tokens, where one classification token [CLS] was used as the first token and several separation tokens [SEP] were used as separators of different information, as follows:

In this work, we first used a pretrained BERT as the encoder. The purpose was to convert the textual question with marks assigned by schema linking and database schema into hidden representations via the multihead attention mechanism [29]. Then, for each span, we took the average of the hidden representations of word and mark as the span representation. Last, through a fully connected pooling layer, we obtained the final hidden representations of question H_X, column names H_C, and table names H_T. The nonlinear transformation of [CLS] representation $tanh(WH_{CLS}+b)$ was used to initialize the decoder, where W and b were trainable parameters.

Tree-Structured Intermediate Representation

Tree-structured intermediate representation refers to a syntax tree that bridges the question and the SQL query. It contains SQL information implicated in the textual question and could be transformed to a SQL query more intuitively due to the nature of the tree structure of SQL. Figure 3 demonstrates an example of generating an intermediate tree from an input question. Different from the previous works that used the grammar to assist the SQL generation [19,22], we designed the grammatical rules based on the MIMICSQL dataset. Specifically, we kept only the necessary rules occurring in the dataset to make the prediction more accurately and added a new rule for predicting the condition value.

Figure 3. Example of tree-structured intermediate representation: (a) grammar rules that transform the SQL query into an abstract syntax tree, (b) example of the action sequence generated by the grammar-based decoder with 4 types of actions, and (c) intermediate tree constructed from the action sequence in b following the grammar rules in a.

$Z \rightarrow R$

$R \rightarrow Select \mid Select\ Filter \mid distinct\ Select \mid distinct\ Select\ Filter$

$Select \rightarrow A \mid A\ A \mid A\ A\ A \mid A\ A\ A\ A \mid A\ A\ A\ A\ A$

$Filter \rightarrow and\ Filter\ Filter \mid or\ Filter\ Filter \mid mt\ A\ V \mid lt\ A\ V \mid mte\ A\ V \mid lte\ A\ V \mid e\ A\ V \mid ne\ A\ V$

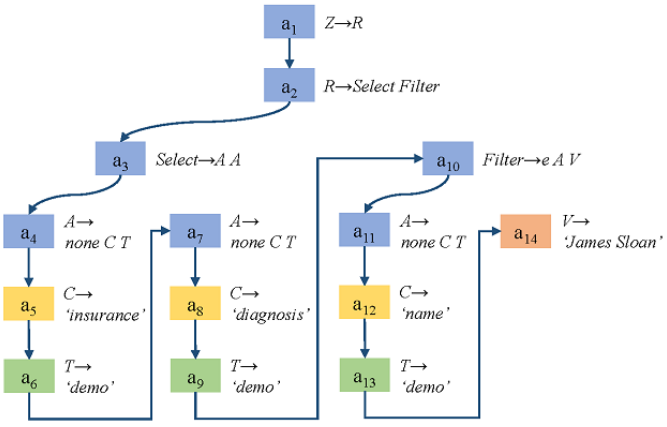
$A \rightarrow max\ C\ T \mid min\ C\ T \mid count\ C\ T \mid sum\ C\ T \mid avg\ C\ T \mid none\ C\ T$

$C \rightarrow [column]$

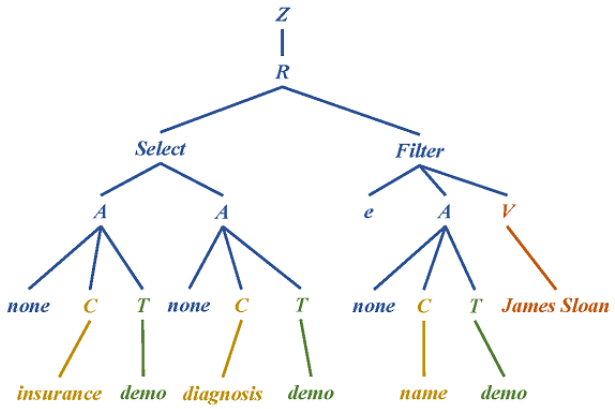
$T \rightarrow [table]$

$V \rightarrow [value]$

(a)



(b)



(c)

To construct the intermediate tree from a SQL query, we first defined a set of grammar rules, as shown in Figure 3a. The intermediate tree starts from a root node *Z*. Since there are no complicated SQL components such as *Union* in this task, a single node *R* was directly attached to *Z*. Then, we attached a node *Select* or *Filter* under *R*, which was determined by the *Select* clause or *Where* clause, respectively. For the subtree of node *Select*, according to the number of columns in the *Select* clause, the same number of nodes *A* were attached to node *Select*. Each node *A* comprised an aggregation function node, a node *C*, and a node *T*. The aggregation function could be either of *none*, *max*, *min*, *count*, etc, while node *C* denoted the column name and node *T* denoted the table name. For the subtree of node *Filter*, it was determined by the conditions in the *Where* clause. If there was more than one condition, the corresponding number of *Filter* nodes would be attached. Next, for each *Filter*

node, it attached a relational operator, a node *A*, and a node *V*. Relational operators include *more than* (*mt*), *less than* (*lt*), *equal* (*e*), etc. Node *V* denotes the condition value. The intermediate tree in Figure 3c was generated by the action sequence in Figure 3b following the grammar rules defined in Figure 3a. The generation process was in the depth-first, left-to-right order.

Grammar-Based Decoder

The generation process of the intermediate tree was formalized into sequential applications of actions. The actions either applied a production rule on the derivation tree or produced a terminal node. According to the grammar rules, we defined 4 types of actions (ie, *ApplyRule*, *SelectColumn*, *SelectTable*, and *SelectValue*) and adopted the grammar-based decoding strategy [30,31]. *ApplyRule*(*r*) applied a production rule *r* to construct the skeleton of the intermediate tree, and the other 3 types of actions were designed to produce the terminal tokens. Thus, the goal of the decoder was to generate an action sequence *A* based on the outputs of the encoder. Formally, the decoding process was formalized as follows:

where a_i was the action taken at time step *i*, $a_{<i}$ was the sequence of actions before *i*, and *n* was the number of total time steps of the whole action sequence.

The probability of selecting a rule *r* as the current action a_i was calculated as follows:

where  denoted the current hidden state of LSTM, v_i and u_i denoted the context vectors that were obtained by performing attention over H_X and $[H_C;H_T]$, $e(r)$ was the one-hot vector for rule *r*.

The *SelectColumn* action was implemented via a memory-enhanced pointer network to select a column *c*, in which the memory was used to record the selected columns [32]. Once a column was selected, it was removed from the schema and recorded in the memory. The probability of selecting a column *c* was calculated as follows:

where SCH denoted selecting from the schema, MEM denoted selecting from memory, and  and  denoted the corresponding hidden representations of columns.

For the *SelectTable* action, we leveraged the relationship between columns and tables to prune irrelevant tables. Thus, the decoder predicted the table *t* that the selected columns belong to. The probability of choosing a table *t* was calculated as follows:

As for *SelectValue*, since the value was always mentioned in the textual question, the decoder extracted a condition value *v* by finding a start position and an end position from the question via 2 different pointer networks, respectively, as follows:

where p_{start} and p_{end} denoted the probabilities of the start and end positions.

Afterward, in order to keep the extracted value consistent with the value in the database, we also adopted the condition-value-recover technique proposed in TREQS [23] to find the most similar value in the look-up table content by computing the ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation based on the Longest Common Subsequence) [33] score between them.

SQL Query Generation

According to the grammar rules in Figure 3, when inferring a SQL query from an intermediate tree, we traversed the whole intermediate tree and mapped each node to the corresponding SQL component. The production rule applied on node *Z* denoted that it was just a single SQL query. The node *R* represented the start point. Following the child nodes of node *R*, we generated the skeleton of a SQL query, such as whether the SQL query had a *Select* clause or *Where* clause corresponding to the node *Select* and *Filter*, respectively. The node *Select* indicated how many columns the *Select* clause had. The node *Filter* indicated how many different conditions were in the *Where* clause. Based on the subtree of node *Select* or *Filter*, we filled in the details (ie, the aggregation function, relational operator, column name, table name, and condition value). The *From* clause was generated from the nodes of selected tables by identifying the shortest path that connected these tables in the schema.

Experimental Settings

We adopted the pretrained uncased base BERT as our encoder, and the hidden size was set as 768. For the decoder module, the hidden size of LSTM was set as 300. The maximum length of the action sequence was set as 128. The size of the attention vector was set as 300. The coarse-to-fine framework [34-36] was used to model the generation process. The Adam optimizer [37] was adopted to train the model parameters for 100 epochs. The learning rate was set as 1e-06, and gradient clipping was used with a maximum gradient norm of 5.0. During training, we set the batch size as 8. The numbers of *ApplyRule*, *SelectColumn*, and *SelectTable* candidate actions were 24, 39, and 5 respectively. The size of the *SelectValue* candidate action was based on the length of the input textual question. We selected the model which achieved the best performance on the validation set. The MedTS was implemented with PyTorch [38] and trained on a Tesla V100 GPU (NVIDIA Corp). Our code has been shared on GitHub to facilitate other researchers [39].

We compared our proposed MedTS with 5 competitor methods. Seq2Seq [14] is an LSTM-based model with the attention mechanism, in which the SQL query is regarded as an ordinary word sequence. PtrGen [15] is a Seq2Seq-based pointer-generator network, which can directly copy the word from the input question to alleviate the repetition and out-of-vocabulary (OOV) phenomenon. SQLNet [18] is a sketch-based text-to-SQL model to avoid the order problems

that occurred in the Seq2Seq model. Coarse2Fine [36] is a 2-stage neural architecture for text-to-SQL. A classifier is first used to obtain a rough sketch of the SQL query and then the details of SQL are filled in based on the input and the sketch individually. TREQS [23] is also a 2-stage text-to-SQL model, including an attentive-copying mechanism and condition value recovery mechanism.

All text-to-SQL methods were evaluated with 2 popular metrics [5], execution accuracy (Acc_{EX}) and logic form accuracy (Acc_{LF}), which are complementary to evaluate the quality of the generation of SQL queries.

- $Acc_{EX}=N_{EX}/N$, where N denotes the total number of question-SQL pairs and N_{EX} denotes the number of SQL queries that can be executed and achieve the correct answers
- $Acc_{LF}=N_{LF}/N$, where N_{LF} denotes the number of queries that match exactly with the ground truth of the SQL query

Results

Quantitative Evaluation

Table 2 provides the quantitative results on the validation and test sets. Seq2Seq achieved 0.103 Acc_{LF} and 0.173 Acc_{EX} on the test set. SQLNet performed better than Seq2Seq, since it considered the dependencies between the components of SQL query based on a graph derived from the sketch. But it was not easy to cover all the queries. PtrGen performed much better than SQLNet with 0.180 Acc_{LF} and 0.292 Acc_{EX} on the test set because it directly extracted words from textual questions to reduce the OOV words, especially when most values occurred in the original question. Coarse2Fine achieved decent performance since it incorporated the schema information into question encoding, but it was limited by the number of sketches and had difficulty handling more complex SQL. TREQS further improved the performance via several effective mechanisms, such as controlled generation and placeholder replacement. But it is just based on the Seq2Seq framework and did not consider the intrinsic structure information of SQL itself. Compared to all the methods mentioned above, MedTS achieved the best performance with 0.681 Acc_{LF} and 0.880 Acc_{EX} on the validation set and 0.784 Acc_{LF} and 0.899 Acc_{EX} on the test set, which outperformed the best competitor method by at least 29% and 27% in terms of Acc_{LF} and Acc_{EX}, respectively, on the test set.

Table 2. The logic form accuracy (Acc_{LF}) and execution accuracy (Acc_{EX}) of SQL query generated by various methods.

Methods	Validation		Test	
	Acc_{LF} [a]	Acc_{EX} [b]	Acc_{LF}	Acc_{EX}
Seq2Seq	0.092	0.195	0.103	0.173
SQLNet	0.086	0.225	0.142	0.260
PtrGen	0.181	0.325	0.180	0.292
Coarse2Fine	0.217	0.309	0.378	0.496
TREQS	0.562	0.675	0.556	0.654
MedTS	0.681	0.880	0.784	0.899

[a]Acc_{LF}: logic form accuracy.

[b]Acc_{EX}: execution accuracy.

Performance on Each Component of SQL

In order to further analyze the generation result, we broke down the SQL queries into 5 components according to the SQL grammar structure, including aggregation operation, aggregation column, table, condition column along with its operation, and condition value. The experimental results are shown in Table 3. Since Coarse2Fine cannot handle multitable questions and is limited by table-aware assumption, its performance cannot be compared to other methods. Aggregation operation refers to the operations in the *Select* clause used to aggregate all the values of a column and return a single value, such as *Count*, *Sum*, *Avg*, etc. All methods except for Coarse2Fine achieved a very high accuracy of more than 97%. Aggregation column was the target column in the *Select* clause for the aggregation operation. MedTS outperformed other methods significantly by at least 5% and 12% on validation and test sets, respectively. Table was the target table in the *From* clause. Except for the Coarse2Fine, the other methods achieved similar accuracy. MedTS achieved the best performance. Condition column along with its operation represented the column and operation in the *Where* clause. Compared to the other competitor methods, MedTS achieved a large improvement by at least 8% on the test set. Condition value refers to the condition value in the *Where* clause. It was observable that the performance on condition value primarily played a vital role in the overall SQL generation performance. MedTS achieved improvement by at least 11% on the test set. In summary, the experimental results of MedTS on each component of SQL were relatively balanced and better, especially the performance on aggregation column and condition value.

Table 3. Accuracy of each component of SQL query.

Methods	Validation					Test				
	Agg [a] *op* [b]	Agg_{col} [c]	*Table*	*Con* [d] *c+o* [e]	Con_{val} [f]	Agg_{op}	Agg_{col}	*Table*	Con_{c+o}	Con_{val}
Coarse2Fine	0.321	0.313	0.321	0.260	0.214	0.524	0.490	0.528	0.448	0.413
Seq2Seq	0.978	0.872	0.926	0.471	0.174	0.970	0.696	0.892	0.565	0.296
SQLNet	0.994	0.939	0.933	0.722	0.080	0.989	0.873	0.941	0.749	0.140
PtrGen	0.987	0.917	0.944	0.795	0.236	0.987	0.830	0.926	0.824	0.235
TREQS	0.990	0.912	0.942	0.834	0.694	0.993	0.827	0.941	0.844	0.763
MedTS	0.994	0.988	0.971	0.893	0.785	0.991	0.985	0.951	0.919	0.851

[a]*Agg*: aggregation.

[b]*Op*: operation.

[c]*Col*: column.

[d]*Con*: condition.

[e]*c+o*: column and operation.

[f]*Val*: value.

Ablation Study

We also conducted an ablation study to analyze the impact of schema linking as well as the use of different types of pretrained representations on question encoding and show the results in Table 4. When the schema linking was not used, the performance of MedTS dropped by 1.4% on Acc_{LF} and 1.3% on Acc_{EX} on the test set, which demonstrated the effectiveness of schema linking. The tested pretrained representations included a recurrent neural network (RNN)-based encoder (ie, BioWord2Vec [40]) and two BERT-based encoders (ie, ClinicalBERT [41] and BioBERT [42]). As shown in Table 4, the RNN-based encoder with pretrained BioWord2Vec performed far worse than the BERT-based encoder by at least 21.0% on Acc_{LF} and 17.9% on Acc_{EX} on the test set. We argue that the main reason is that the LSTM encoder cannot model the interaction of the entire sequence itself. As for the BERT-based encoders, we observed that the performance of ClinicalBERT was inferior to the others since it specializes in clinical notes that are obviously different from the natural language text. Compared to MedTS (with uncased base BERT), BioBERT achieved slightly better performance since it uses the medical literature for pretraining which is more beneficial to the representations of medical questions.

Table 4. The experimental results of the ablation study.

Methods	Validation		Test	
	Acc_{LF} [a]	Acc_{EX} [b]	Acc_{LF}	Acc_{EX}
MedTS	0.681	0.880	0.784	0.899
w/o SL	0.669	0.870	0.773	0.887
w/ BioWord2Vec	0.472	0.690	0.501	0.644
w/ ClinicalBERT	0.556	0.771	0.634	0.784
w/ BioBERT	0.684	0.882	0.790	0.904

[a]Acc_{LF}: logic form accuracy.

[b]Acc_{EX}: execution accuracy.

Discussion

Principal Findings

Our proposed model MedTS achieved the best Acc_{LF} and Acc_{EX} on the validation and test sets, with pretrained encoder and grammar-based decoder. The abstract syntax tree was introduced as the intermediate representation to bridge the gap between medical text and the SQL query. The primary outcomes of this study were (1) a new state-of-the-art model for medical text-to-SQL generation task was proposed and validated and (2) an online demonstration system with the capabilities of transforming the medical text to SQL query and further returning the query results was provided. Experimental results demonstrated that MedTS has great potential to help medical experts facilitate clinical information retrieval and improve the efficiency of decision-making for medical diagnosis and treatment.

Model Performance

MedTS has the ability to capture the semantic relationship between words within textual questions and the dependency relationship between the text and database schema, benefitting from the multihead attention mechanism adopted by the

pretrained encoder. It is difficult for competitor methods to obtain information as rich using the RNN-based encoder. Meanwhile, MedTS can effectively reduce the search space via the grammar-based decoding strategy, which predefines grammatical rules and introduces the tree-structured intermediate representation. Although several mechanisms were designed in the competitor methods to make the generated SQL query more accurate, they still view the SQL query as an ordinary word sequence and ignore the intrinsic structure characteristic of SQL itself, which makes them perform worse than MedTS.

Case Study

In addition to quantitative evaluations, we conducted an extensive set of qualitative case studies on the test data to analyze the generated SQL query. We manually analyzed all 1000 text-SQL pairs in the test set. Among them, 784 generated SQL queries that were entirely consistent with the ground truth, and 115 generated SQL queries that were not identical to the ground truth in the logical form but also achieved accurate execution results. Most of them were caused by the different positions of the 2 tables connected by the *join* operation (eg, example 1 in Table 5). This phenomenon also explains why the quantitative evaluation results of Acc_{EX} are higher than Acc_{LF} in Table 2. In addition, 5 generated SQL queries were correct but considered wrong by the Acc_{EX} because of the various orders of column in the *select* clause (eg, example 2 in Table 5). The remaining 96 pairs generated incorrect SQL queries. We grouped their errors into different categories from 2 perspectives: clause and element. The clauses included *select*, *join*, and *where*, and the elements included operator, table, column, value, and others. The statistical results are shown in Table 6. Note that there was no operator in the *join* clause. Similarly, since the value only presented in the *where* clause, the value error of *select* and *join* clauses was none. When there was a table error in the *where* clause, it was usually due to the wrong decision in the *select* or *join* clauses, so we did not count these types of errors again. The rest of the errors, such as more or less conditions, are grouped into other categories.

From the element's perspective, we observed that the prediction errors of *column* and *value* account for the majority. From the perspective of the clause, more than 50% of clause errors were in *where* clauses, while most *where* clause errors were due to incorrect values or columns. Example 3 in Table 5 is a representative case of *where* clause error due to the incorrect value. The value of *expire_flag* is a numeric type in SQL but a text description in the question. Example 4 in Table 5 shows a case of *where* clause error due to the wrong column, in which the *admityear* and *dob_year* are semantically close, leading to the wrong choice. It was challenging to achieve high accuracy in these cases, since MedTS is based on the pointer network that selects terms from textual questions to generate SQL queries. The *operation* error means that the condition column and value in the *where* clause are correct but the operator is wrong, which may return completely opposite results, as shown by example 5 in Table 5.

Table 5. Five representative examples of qualitative case study.

Examples
Example 1
Q[a]: Let me know the short title and ICD-9[b] codes of diagnoses for patient John Gartman.
G[c]: Select diagnoses.“icd9_code,” diagnoses.“short_title” from demographic inner join diagnoses on demographic.hadm_id = diagnoses.hadm_id where demographic.“name” = “john gartman”
P[d]: Select diagnoses.“icd9_code,” diagnoses.“short_title” from diagnoses inner join demographic on diagnoses.hadm_id = demographic.hadm_id where demographic.“name” = “john gartman”
Example 2
Q: Tell me which primary disease the patient Walter Locher is suffering from and whether he is still alive or not.
G: Select demographic.“expire_flag,” demographic.“diagnosis” from demographic where demographic.“name” = “walter locher”
P: Select demographic.“diagnosis,” demographic.“expire_flag” from demographic where demographic.“name” = “walter locher”
Example 3
Q: Calculate the number of dead patients who were admitted to hospital before 2123.
G: Select count (distinct demographic.“subject_id”) from demographic where demographic.“expire_flag” = “1” and demographic.“admityear” < “2123”
P: Select count (distinct demographic.“subject_id”) from demographic where demographic.“expire_flag” = “0” and demographic.“admityear” < “2123”
Example 4
Q: How many American Indian/Alaska Native ethnic background patients were born before 2148?
G: Select count (distinct demographic.“subject_id”) from demographic where demographic.“ethnicity” = “american indian/alaska native” and demographic.“admityear” < “2148”
P: Select count (distinct demographic.“subject_id”) from demographic where demographic.“ethnicity” = “american indian/alaska native” and demographic.“dob_year” < “2184”
Example 5
Q: Find the minimum number of days of hospital stay for patients born before the year 2200.
G: Select min (demographic.“days_stay”) from demographic where demographic.“dob_year” > “2200”
P: Select min (demographic.“days_stay”) from demographic where demographic.“dob_year” < “2200”

[a]Q: textual question.

[b]ICD-9: International Classification of Diseases Clinical Modification, 9th Revision.

[c]G: golden truth.

[d]P: predicted result.

Table 6. Statistical analysis of error categories.

	Select	Join	Where	#Element Error (%)
Operator, n	9	—[a]	3	12 (10.6)
Table, n	8	6	—	14 (12.4)
Column, n	17	—	10	27 (23.9)
Value, n	—	—	44	44 (38.9)
Other, n	4	9	3	16 (14.2)
#Clause Error (%)	38 (33.6)	15 (13.3)	60 (53.1)	113 (100)

[a]Not applicable.

Comparison With Prior Work

In the medical field, a few studies have focused on the text-to-SQL task, but most of them either proposed rule-based methods [20,21] or validated on the small-scale datasets [22]. Wang et al [23] constructed the first large-scale medical text–to-SQL dataset and proposed a neural model TREQS to undertake this task. However, TREQS focused on solving the OOV problem and condition value generation. Compared with the rule-based methods, our proposed model has better

applicability and can be extended to other datasets. Compared with the previous neural models, our model adapts more advanced deep learning methods to this task and achieves the optimal experimental performance on a large-scale dataset.

Limitations and Future Work

As discussed above, several problems are still to be solved, such as improving the accuracy of the *conditioncolumn* and *value* in the *where* clause, especially the gap between natural language description and the value stored in the database. In future work, we will continue to improve the accuracy and robustness of the model (eg, introducing more schema information such as the data type of column to achieve the goal of practical deployment). In addition, the form of question and SQL in MIMICSQL is relatively simple, which is not enough to cover various situations in the practical applications. Therefore, we plan to keep exploring different data forms for more practical scenarios, such as generating SQL queries containing more complex clauses.

Conclusion

In this work, we proposed a medical text–to-SQL method named MedTS, which incorporates a BERT-based attention encoder to obtain schema-enhanced text representation and a grammar-based LSTM decoder to generate the intermediate action sequence before generating a SQL query. By introducing the intermediate representation, MedTS can reduce the search space during decoding and mitigate the mismatch problem between the medical question and the SQL query. Experiments on the MIMICSQL dataset demonstrate that MedTS substantially outperforms the state-of-the-art methods. Further analyses on each component of SQL query and the case study confirm MedTS's effectiveness and robustness, demonstrating its strong potential.

Acknowledgments

This work was jointly supported by grants 62006061, 61872113, 62106115, and 62102118 from the Natural Science Foundation of China; grants JCYJ20190806112210067, JCYJ20200109113403826, and JCYJ20200109113441941 from the Strategic Emerging Industry Development Special Funds of Shenzhen; grant CCF-BAIDUOF2020004 from the CCF-Baidu Open Fund; and grant GXWD20201230155427003-20200824155011001 from the Stable Support Program for Higher Education Institutions of Shenzhen.

Authors' Contributions

YP and CW proposed the methods, designed and performed the experiments, and drafted the manuscript. BH supervised the research and participated in the study design. YX critically revised the manuscript and made substantial contributions to interpreting the results. XW and QC provided guidance and reviewed the manuscript. JC and JD participated in the manuscript review. All authors provided feedback and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Shao S, Chan Y, Kao Yang Y, Lin S, Hung M, Chien R, et al. The Chang Gung Research Database: a multi-institutional electronic medical records database for real-world epidemiological studies in Taiwan. Pharmacoepidemiol Drug Saf 2019 May;28(5):593-600. [doi: 10.1002/pds.4713] [Medline: 30648314]
2. Garies S, Birtwhistle R, Drummond N, Queenan J, Williamson T. Data resource profile: national electronic medical record data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Int J Epidemiol 2017 Aug 01;46(4):1091-1092. [doi: 10.1093/ije/dyw248] [Medline: 28338877]
3. Garies S, Cummings M, Quan H, McBrien K, Drummond N, Manca D, et al. Methods to improve the quality of smoking records in a primary care EMR database: exploring multiple imputation and pattern-matching algorithms. BMC Med Inform Decis Mak 2020 Mar 14;20(1):56 [FREE Full text] [doi: 10.1186/s12911-020-1068-5] [Medline: 32171301]
4. Yaghmazadeh N, Wang Y, Dillig I, Dillig T. SQLizer: query synthesis from natural language. Proc ACM Program Lang 2017 Oct 12;1(OOPSLA):1-26. [doi: 10.1145/3133887]
5. Zhong V, Xiong C, Socher R. Seq2sql: generating structured queries from natural language using reinforcement learning. ArXiv. Preprint posted online on August 30, 2017. [FREE Full text]
6. Price P. Evaluation of spoken language systems: the ATIS domain. Proc Workshop Speech Natural Lang 1990 Jun:91-95. [doi: 10.3115/116580.116612]
7. Dahl D, Bates M, Brown M. Expanding the scope of the ATIS task: the ATIS-3 corpus. Proc Workshop Human Lang Technol 1994:43-48. [doi: 10.3115/1075812.1075823]
8. Iyer S, Konstas I, Cheung A. Learning a neural semantic parser from user feedback. 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; -Aug 4; Vancouver, Canada; 2017 Jul 30; Vancouver. [doi: 10.18653/v1/p17-1089]

9. Zelle J, Mooney R. Learning to parse database queries using inductive logic programming. 1996 Presented at: The Thirteenth National Conference on Artificial Intelligence; 1996; Portland URL: https://www.cs.utexas.edu/~ml/papers/chill-aaai-96.pdf
10. Finegan-Dollak C, Kummerfeld J, Zhang L. Improving text-to-SQL evaluation methodology. 2018 Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; Jul ; Melbourne, Australia; 2018; Melbourne. [doi: 10.18653/v1/p18-1033]
11. Yu T, Zhang R, Yang K. Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels. [doi: 10.18653/v1/d18-1425]
12. Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. Adv Neural Inf Process Syst 2014:3104-3112 [FREE Full text]
13. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. 2015 Presented at: The 3rd International Conference on Learning Representations; 2015; San Diego.
14. Luong M, Pham H, Manning C. Effective approaches to attention-based neural machine translation. 2015 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon. [doi: 10.18653/v1/d15-1166]
15. See A, Liu P, Manning C. Get to the point: summarization with pointer-generator networks. Get to the point: summarization with pointer-generator networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017; Vancouver. [doi: 10.18653/v1/p17-1099]
16. Vinyals O, Kaiser L, Koo T. Grammar as a foreign language. Adv Neural Inf Process Syst. 2015. URL: https://proceedings.neurips.cc/paper/2015/file/277281aada22045c03945dcb2ca6f2ec-Paper.pdf [accessed 2021-11-11]
17. Dong L, Lapata M. Language to logical form with neural attention. 2016 Presented at: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; Aug ; Berlin, Germany; 2016; Berlin. [doi: 10.18653/v1/p16-1004]
18. Xu X, Liu C, Song D. SQLnet: generating structured queries from natural language without reinforcement learning. ArXiv. Preprint posted online on November 13, 2017. [FREE Full text]
19. Guo J, Zhan Z, Gao Y. Towards complex text-to-SQL in cross-domain database with intermediate representation. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence. [doi: 10.18653/v1/p19-1444]
20. Ben AA, Zweigenbaum P. Medical question answering: translating medical questions into SPARQL queries. 2012 Presented at: Proceedings of the 2nd ACM SIGHIT international health informatics symposium; 2012; Miami. [doi: 10.1145/2110363.2110372]
21. Roberts K, Patra BG. A semantic parsing method for mapping clinical questions to logical forms. AMIA Annu Symp Proc 2017;2017:1478-1487 [FREE Full text] [Medline: 29854217]
22. Yu X, Chen T, Yu Z. Dataset and enhanced model for eligibility criteria-to-SQL semantic parsing. 2020 Presented at: The 12th International Conference on Language Resources and Evaluation; 2020; Marseille.
23. Wang P, Shi T, Reddy C. Text-to-SQL generation for question answering on electronic medical records. 2020 Presented at: Proceedings of The Web Conference; 2020; Taipei. [doi: 10.1145/3366423.3380120]
24. 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;3:160035 [FREE Full text] [doi: 10.1038/sdata.2016.35] [Medline: 27219127]
25. Devlin J, Chang M, Lee K. Bert: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics; 2019; Minneapolis.
26. 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]
27. MedTS: a BERT-based generation model to transform medical texts to SQL queries for electronic medical records. URL: http://112.74.48.115:9201/ [accessed 2021-11-06]
28. Pampari A, Raghavan P, Liang J. emrQA: a large corpus for question answering on electronic medical records. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels. [doi: 10.18653/v1/d18-1258]
29. Vaswani A, Shazeer N, Parmar N. Attention is all you need. ArXiv. Preprint posted online on June 12, 2017. [FREE Full text]
30. Yin P, Neubig G. A syntactic neural model for general-purpose code generation. 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017; Vancouver. [doi: 10.18653/v1/p17-1041]
31. Yin P, Neubig G. A transition-based neural abstract syntax parser for semantic parsing and code generation. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels. [doi: 10.18653/v1/d18-2002]

32. Liang C, Berant J. Neural symbolic machines: learning semantic parsers on freebase with weak supervision. 2017 Presented at: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics; 2017; Vancouver. [doi: 10.18653/v1/p17-1003]
33. Lin C, Hovy E. Manual and automatic evaluation of summaries. 2002 Presented at: Proceedings of the ACL-02 Workshop on Automatic Summarization; 2002; Phildadelphia. [doi: 10.3115/1118162.1118168]
34. Solar-Lezama A. Program synthesis by sketching [Thesis]. Berkeley: University of California, Berkeley; Dec 18, 2008.
35. Bornholt J, Torlak E, Grossman D. Optimizing synthesis with metasketches. 2016 Presented at: Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages; 2016; St. Petersburg. [doi: 10.1145/2837614.2837666]
36. Dong L, Lapata M. Coarse-to-fine decoding for neural semantic parsing. 2018 Presented at: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; 2018; Melbourne. [doi: 10.18653/v1/p18-1068]
37. Kingma D, Ba J. Adam: a method for stochastic optimization. 2015 Presented at: The 3rd International Conference on Learning Representations; 2015; San Diego.
38. Paszke A, Gross S, Massa F. Pytorch: an imperative style, high-performance deep learning library. ArXiv. Preprint posted online on December 3, 2019. [FREE Full text]
39. Code at GitHub: pan915/MedTS. URL: https://github.com/pan915/MedTS [accessed 2021-11-06]
40. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec: improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):1-9 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
41. Huang K, Altosaar J, Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission. ArXiv. Preprint posted online on April 10, 2019. [FREE Full text] [doi: 10.1090/mbk/121/79]
42. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]

Abbreviations

Acc_{EX}: execution accuracy
Acc_{LF}: logic form accuracy
BERT: Bidirectional Encoder Representations from Transformers
EMR: electronic medical record
LSTM: long short-term memory
MedTS: medical text–to-SQL
MIMIC III: Medical Information Mart for Intensive Care III
OOV: out-of-vocabulary
RNN: recurrent neural network
Seq2Seq: sequence-to-sequence
SPARQL: SPARQL Protocol and RDF Query Language
TREQS: Translate-Edit Model for Question-to-SQL
ROUGE-L: Recall-Oriented Understudy for Gisting Evaluation based on the Longest Common Subsequence

Edited by G Eysenbach; submitted 06.08.21; peer-reviewed by Y Kim, Q Jia; comments to author 27.08.21; revised version received 23.10.21; accepted 27.10.21; published 08.12.21.

Please cite as:
Pan Y, Wang C, Hu B, Xiang Y, Wang X, Chen Q, Chen J, Du J
A BERT-Based Generation Model to Transform Medical Texts to SQL Queries for Electronic Medical Records: Model Development and Validation
JMIR Med Inform 2021;9(12):e32698
URL: https://medinform.jmir.org/2021/12/e32698
doi:10.2196/32698
PMID:34889749

Original Paper

A Smartphone App (AnSim) With Various Types and Forms of Messages Using the Transtheoretical Model for Cardiac Rehabilitation in Patients With Coronary Artery Disease: Development and Usability Study

Jah Yeon Choi[1*], MD, PhD; Ji Bak Kim[1*], MD, PhD; Sunki Lee[2], MD; Seo-Joon Lee[3], PhD; Seung Eon Shin[1], BS; Se Hyun Park[4], PhD; Eun Jin Park[1], MD, PhD; Woohyeun Kim[1], MD, PhD; Jin Oh Na[1], MD, PhD; Cheol Ung Choi[1], MD, PhD; Seung-Woon Rha[1], MD, PhD; Chang Gyu Park[1], MD, PhD; Hong Seog Seo[1], MD, PhD; Jeonghoon Ahn[5], PhD; Hyun-Ghang Jeong[6], MD, PhD; Eung Ju Kim[4], MD, PhD

[1]Cardiovascular Center, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[2]Division of Cardiology, Department of Internal Medicine, Hallym University Dongtan Sacred Heart Hospital, Dongtan, Republic of Korea

[3]Department of Medical Informatics, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[4]Sports Medical Center, Seoul, Republic of Korea

[5]Department of Health Convergence, Ewha Womans University, Seoul, Republic of Korea

[6]Department of Psychiatry, Korea University Guro Hospital, Korea University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Eung Ju Kim, MD, PhD
Sports Medical Center
Korea University Guro Hospital
Seoul, 08308
Republic of Korea
Phone: 82 2 2626 3020
Fax: 82 2 864 3062
Email: withnoel@empas.com

Abstract

Background: Despite strong evidence of clinical benefit, cardiac rehabilitation (CR) programs are currently underutilized and smartphone-based CR strategies are thought to address this unmet need. However, data regarding the detailed process of development are scarce.

Objective: This study focused on the development of a smartphone-based, patient-specific, messaging app for patients who have undergone percutaneous coronary intervention (PCI).

Methods: The AnSim app was developed in collaboration with a multidisciplinary team that included cardiologists, psychiatrists, nurses, pharmacists, nutritionists, and rehabilitation doctors and therapists. First, a focus group interview was conducted, and the narratives of the patients were analyzed to identify their needs and preferences. Based on the results, health care experts and clinicians drafted messages into 5 categories: (1) general information regarding cardiovascular health and medications, (2) nutrition, (3) physical activity, (4) destressing, and (5) smoking cessation. In each category, 90 messages were developed according to 3 simplified steps of the transtheoretical model of behavioral change: (1) precontemplation, (2) contemplation and preparation, and (3) action and maintenance. After an internal review and feedback from potential users, a bank of 450 messages was developed.

Results: The focus interview was conducted with 8 patients with PCI within 1 year, and 450 messages, including various forms of multimedia, were developed based on the transtheoretical model of behavioral change in each category. Positive feedback was obtained from the potential users (n=458). The mean Likert scale score was 3.95 (SD 0.39) and 3.91 (SD 0.39) for readability and usefulness, respectively, and several messages were refined based on the feedback. Finally, the patient-specific message delivery system was developed according to the baseline characteristics and stages of behavioral change in each participant.

Conclusions: We developed an app (AnSim), which includes a bank of 450 patient-specific messages, that provides various medical information and CR programs regarding coronary heart disease. The detailed process of multidisciplinary collaboration

over the course of the study provides a scientific basis for various medical professionals planning smartphone-based clinical research.

(JMIR Med Inform 2021;9(12):e23285) doi:10.2196/23285

KEYWORDS

cardiac rehabilitation; smartphone app; coronary heart disease

Introduction

Coronary heart disease (CHD) is a major cause of death [1,2], especially in developing countries [3]. Over the last few decades, there have been many advances in cardiovascular treatment and treatment strategies in patients with atherosclerotic cardiovascular disease, but a residual risk for recurrent cardiovascular events still exists [4]. Various treatment strategies have emerged for secondary prevention, such as optimal medical therapy, including high doses of statins or proprotein convertase subtilisin/kexin type 9 (PCSK9) inhibitors. However, relatively little attention has been paid to lifestyle modification and cardiac rehabilitation (CR).

CR programs deliver comprehensive clinical information, patient support, and monitor patient status. Recent studies have consistently reported the clinical benefits of CR, such as improved survival, reduction of hospital admissions, and improvements in the quality of life [5-7]. Current guidelines strongly recommend CR for secondary prevention [8-10]. However, CR is so underutilized that the participation rate after acute coronary syndrome or revascularization is only 20%-50% [5]. Furthermore, the adherence rate to CR programs at 6 months was only one-third [11]. Although a low referral rate to CR is one of the main factors related to poor participation or adherence, there are several other factors, such as old age, female sex, geographic distance, low physical activity, costs, and lack of insurance coverage, which are difficult or impossible to change [11-14]. Thus, a new model for enhancing the delivery and maintenance of CR services in patients with CHD is required to improve clinical outcomes and reduce social costs.

Recently, several studies have proved the effectiveness of SMS text messages, a simpler form of intervention compared with hospital-based CR, in improving risk factors and patient adherence to treatment [15-17]. Moreover, smartphone apps are expected to be useful tools for CR as they can deliver various forms of content as well as SMS text messages, and in small studies, they have shown favorable clinical results [18,19]. However, previous text messaging systems have many common limitations as follows: (1) The messaging interventions of the previous studies were primarily in a 1-way direction; therefore, the interaction between the patients and medical experts was limited [15]. (2) Although psychosocial factors influence behavioral change, and psychosocial theory–based programs such as transtheoretical model intervention have shown promising results in patients with cardiovascular disease [20,21], it has not been considered during the app development [14]. (3) The majority of messages are text based [22,23], which might have limitations in education and rehabilitation.

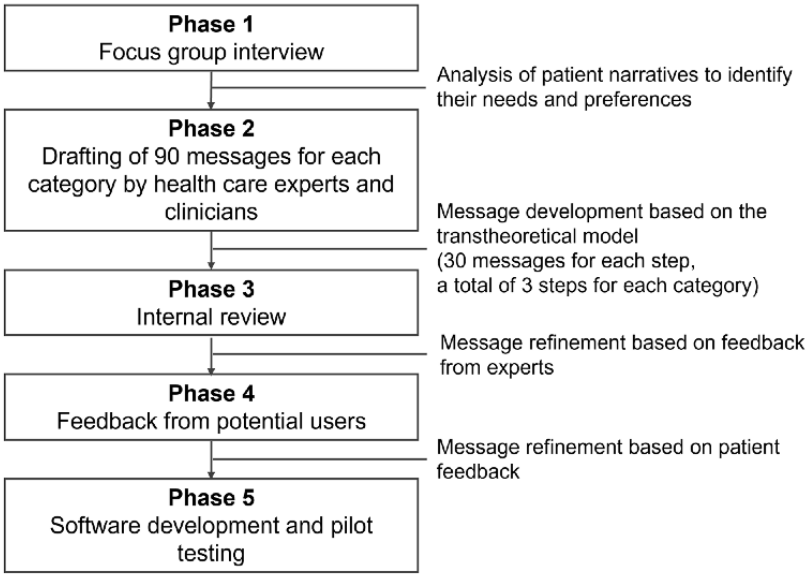
This study focused on developing the Application for Self-improvement (AnSim), a smartphone-based, patient-specific messaging app for patients who have undergone percutaneous coronary intervention (PCI), using the transtheoretical model of behavioral change.

Methods

Process of Message Development

A bank of 450 messages was developed by a multidisciplinary team of cardiologists, psychiatrists, nurses, pharmacists, nutritionists, and rehabilitation doctors and therapists using a 5-phase systematic approach. The scheme of the message development process is illustrated in Figure 1.

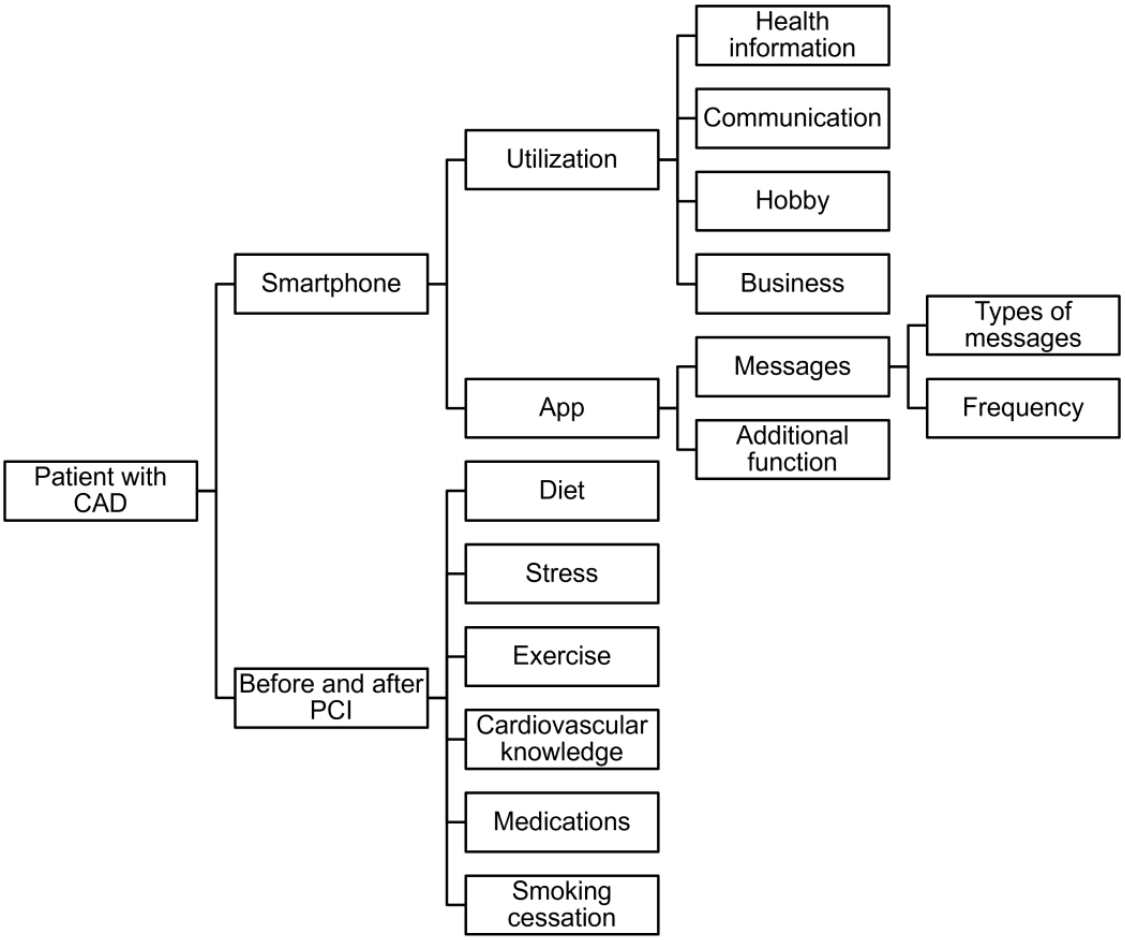
Figure 1. The scheme of the message development process.



Phase 1: Focus Group Interview

A focus group interview was conducted to develop an app that reflects the needs of patients and an understanding of CHD and CR. Patients who had a smartphone and who had undergone PCI within the past 1 year at the Korea University Guro Hospital volunteered for the focus interview. Eight patients of different ages, sexes, and education levels were selected, and in-depth interviews were conducted. The subject of the interview consisted of 5 categories: (1) the degree of smartphone app utilization, (2) exercise, (3) nutrition, (4) stress management, (5) and knowledge about CHD and prevention. The participants of the focus group interview were selected according to the field of CR [24,25] and the design of other CR studies using mobile phone in patients with CHD [15,26,27]. Interviews were recorded with the consent of patients and were transcribed verbatim. Documented interview content was reviewed to exclude repetitive or irrelevant content, such as self-introduction, research participation fee for patients, and personal content to naturally elicit patient's response. Then, 10 nodes (taking medicine, disease, smoking cessation, first diagnosis, recurrence, nutrition, stress, exercise, smartphone, and app) were derived based on the refined interview contents and a conceptual framework that is widely used for qualitative analysis [28,29] (Figure 2). The interview data were coded under each node and analyzed using NVivo (QSR International), a software package for organizing the analysis of qualitative research.

Figure 2. The conceptual framework for the coding. CAD: coronary artery disease; PCI: percutaneous coronary intervention.



Phase 2: Message Development and Its Theoretical Basis

Based on the needs of patients, derived from the focus group interview, experienced health care experts and clinicians drafted messages for the 5 categories: (1) general information regarding cardiovascular health and medications, (2) nutrition, (3) physical activity, (4) destressing, and (5) smoking cessation. Each message was between 40 and 140 Korean characters, in keeping with international guidelines and official educational resources from cardiovascular health–related academic societies.

Furthermore, each message was developed according to 26 behavioral change techniques, which have theoretical backgrounds, such as the information–motivation–behavioral skills model, Theory of Reasoned Action, Theory of Planned Behavior, Social Cognitive Theory, Control Theory, and operant conditioning [30]. In particular, techniques such as the demonstration of behavior and planning of social support, which are difficult to implement with general messaging services, could also be included by providing multimedia, such as sample exercise videos and dietary regimens, and helping participants to connect with smoking cessation centers by actively utilizing smartphone functions. In addition, negative statements in messages were avoided because positive statements are known to help sustainable changes in behavior [31].

To allow participants to receive messages tailored to their stage of behavioral change on a specific topic, messages were developed on the basis of the transtheoretical model of behavioral change, which originally consisted of 5 stages: precontemplation, contemplation, preparation, action, and maintenance [32]. In this study, we simplified these 5 steps into 3 steps: (1) precontemplation, (2) contemplation and preparation, and (3) action and maintenance. According to these 3 simplified steps, behavioral change techniques were categorized, and 30 messages were developed for each step in each category. Finally, a bank of 450 messages was created covering the 5 categories and 3 stages of behavioral change.

Phases 3 and 4: Internal Review and Feedback From Potential Users

The initially developed messages were checked for evidence, appropriateness, and readability, and then the messages were amended through internal review by experts. Each message was corrected or deleted according to the rating (suitable, need to be corrected, unsuitable) given by 9 researchers during an interdepartmental, internal review process. In the next step, feedback was obtained from potential users (n=458) who were outpatients of various ages, sex, and comorbidities with their CHD being treated at the cardiovascular centers of a secondary general hospital (Sejong General Hospital) and a large tertiary general hospital (Korea University Guro Hospital). Each person evaluated 10 messages and rated the readability and usefulness of each message using a 5-point Likert scale survey questionnaire. Simultaneously, free comments were requested for each message. The messages were reviewed again and

refined, based on the feedback from the survey (Multimedia Appendix 1).

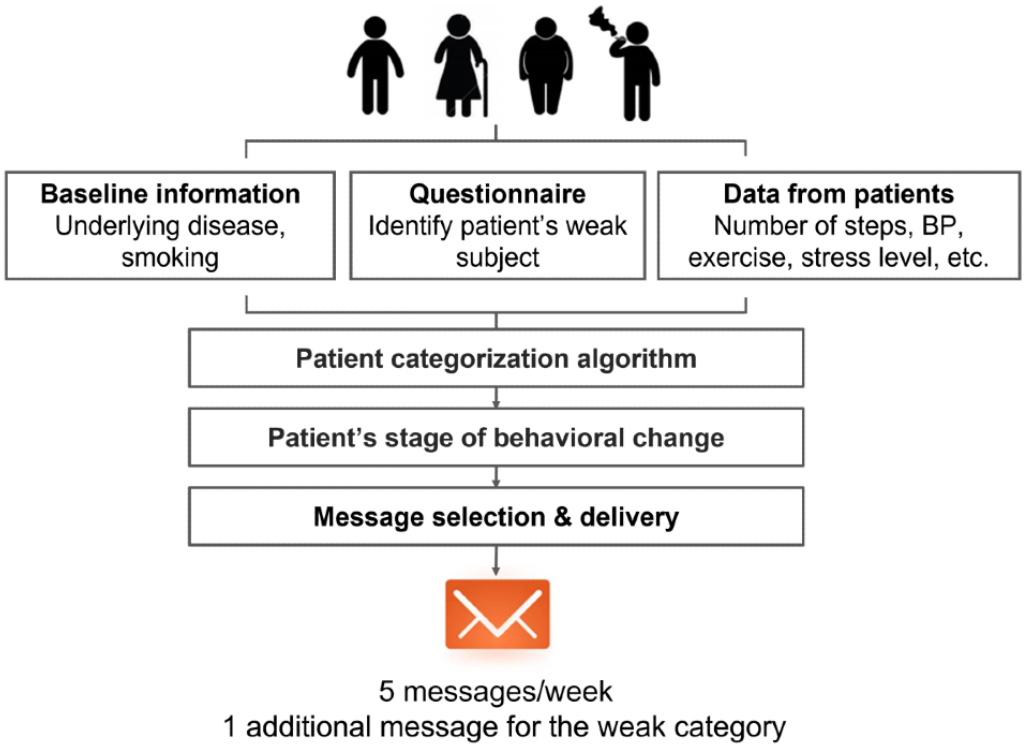
Phase 5: Development of App and Pilot Testing for Message Delivery

Researchers, web designers, and engineers collaborated and developed the user-friendly app by repeatedly discussing the screen composition, function, design, etc. We recruited 20 volunteers to pilot test the message delivery system. All participants were offered brief training at enrollment on how to use the AnSim app and how to input and monitor their health data through the app. Each participant received 6 messages per week for 4 weeks. This frequency was proven to be acceptable for recipients through focus group interviews. Messages were sent randomly at 9 AM, noon, or 3 PM from Monday to Saturday. One message from each of the 5 categories was delivered from the message bank. An additional message was sent for the weak category of each patient.

To provide patient-tailored messaging intervention, baseline characteristics (ie, having diabetes or not, smoking status) of each participant were identified through the enrollment survey, and messages for the dedicated category were selected randomly by an automated system. For example, participants who were nonsmokers or did not have diabetes did not need to receive messages regarding smoking cessation or diabetes management, respectively. Furthermore, each participant's behavior stage was identified by administering simple questionnaires each week (Multimedia Appendix 2), and the messages corresponding to a specific stage of behavior change were sent, and no message was repeated (Figure 3). If the message contained video or audio data that could incur additional data costs, a pop-up message preceded the message saying it must be opened in a Wi-Fi environment.

The number of steps taken in a day was automatically recorded by the AnSim app, and the blood pressure, blood glucose, exercise, diet, stress level, and medicine intake were directly recorded by the participants, although it was not enforced. Instead, to enhance patient participation and motivation, a brief weekly review of health data and support messages was sent every week to participants by an independently designated health care provider from an outsourced health care coaching company.

Figure 3. Delivery of patient-specific messages according to baseline characteristics and stage of behavioral change. BP: blood pressure.



Results

Phase 1 and 2: Focus Group Interview and Message Development

The focus group consisted of 8 patients of different ages, sexes, and education levels who had a smartphone and had undergone PCI within the past 1 year. Detailed patient characteristics are presented in Multimedia Appendix 3. In-depth interviews were conducted, and a summary of the results is presented in Textbox 1. Based on the focus group interview, 90 messages in each of the 5 categories were collected: (1) general information regarding cardiovascular health and medications, (2) nutrition, (3) physical activity, (4) destressing, and (5) smoking cessation. Each message was tailored according to 3 stages of behavioral change: (1) precontemplation, (2) contemplation and preparation, and (3) action and maintenance (Table 1).

Textbox 1. Summary of the focus group interview.

Utilization of smartphone

- Participants had no problem in reading and sending SMS text messages regardless of age.
- Participants below 50 years used smartphones not only for communication, but also for information and business. By contrast, participants over 50 years used smartphones for communication only.
- Positive response for receiving messages on cardiac health.

Exercise

- Concerns about lacking knowledge about proper exercise.

Nutrition

- Participants wanted to know foods and recipes that are good for cardiovascular health.
- Participants tried to avoid fatty foods and eat vegetable-rich diets.

Stress management

- Most of the participants did not know about specific and active stress management method.

Knowledge about coronary artery disease

- Lack of insight regarding recurrence.
- Lack of knowledge about how to prevent recurrence.

Table 1. Examples of messages developed for smoking cessation according to the transtheoretical model of behavior change.

Stage of the transtheoretical model of behavior change	Content	Example (English translation. The original messages were in Korean)
Precontemplation	Provide information about behavior–health link (information–motivation–behavioral skills model)	Smoking is a drug addiction disease, which is registered in the international disease classification.
Contemplation and preparation	Plan social support or social change (social support theories)	Let <NAME>'s family, friends, and co-workers know that you are being treated for heart disease and will quit smoking. In particular, let the friends who smoke know you have heart disease. Everyone will help <NAME> quit smoking.
Action and maintenance	Relapse prevention (Relapse Prevention Theory)	Smoking 1 or 2 cigarettes does not mean that you have failed to quit smoking. Think about the situation in which you smoked, and how you can avoid that particular situation. It may help you in giving up smoking in future.

Phases 3 and 4: Message Refinement

Message feedback was obtained from potential users (n=458) using a 5-point Likert scale survey questionnaire. Each person evaluated 10 different messages and replied about the readability and usefulness, and approximately 9 comments (feedback) were obtained regarding readability (9.47 responses) and usefulness (9.30 responses) in each message. Nearly 98% of messages received more than 3.0 points regarding readability and usefulness, and the average 5-point Likert scale score was 3.95 and 3.91, respectively (Figure 4). Messages with scores less than 3.5 points were further refined and the final expression was formulated by a linguist. Examples of messages developed after refinement are listed in Table 2.

Figure 4. Five point Likert scale scores of various message categories and stages of behavior change.

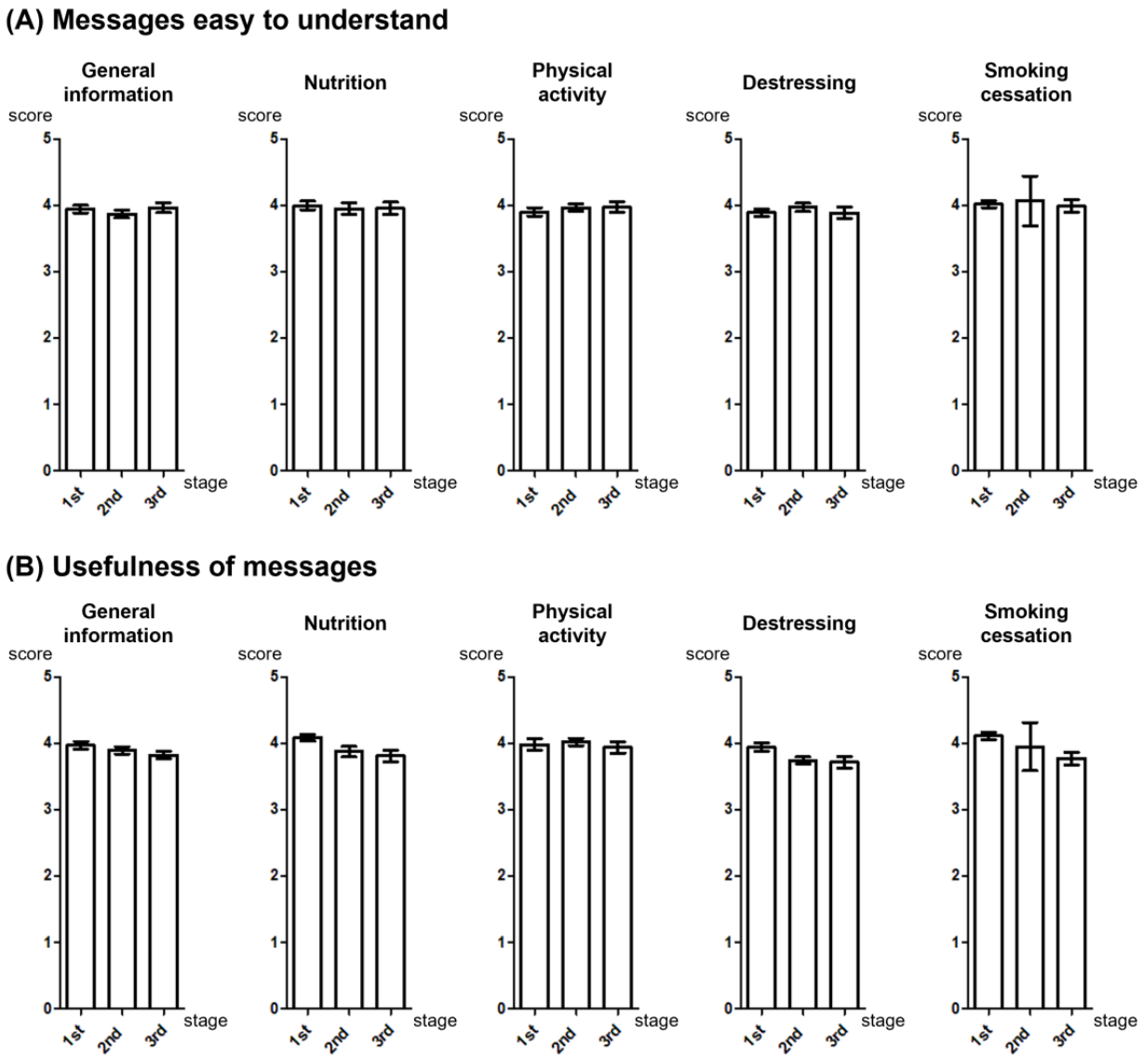


Table 2. Examples of the final version of developed messages after refinement.

Category	Stage of behavior change[a]	Example (English translation. The original messages were in Korean)
General cardiovascular health and medications	2	Taking antiplatelet agents such as aspirin is very important for patients who received percutaneous coronary intervention. When you plan a tooth extraction or endoscopic examination, do not arbitrarily stop the medication without consulting your doctor first.
Nutrition	1	Eating too much salt may burden your heart, leading to swelling and raising the blood pressure (low-salt diet recipes: Link)
Physical activity	3	We applaud you for maintaining a steady routine of exercise. Exercise not only helps you lose weight but also strengthens your heart.
Destress	1	Did you know that depression and coronary artery disease are correlated? Coronary artery disease may lead to depression, and depression in turn can also increase the risk of coronary artery disease.
Smoking cessation	1	Get free smoking cessation counseling. You can get free personalized 1:1 counseling at any time, at home or at work. (Phone number of the national antismoking organization)

[a]1=precontemplation; 2=contemplation and preparation; 3=action and maintenance.

Phase 5: Development of App and Test of the Delivery System

We developed the AnSim app for both Android and iPhone OS versions. Various multimedia forms, such as exercise videos and dietary regimens, and links for smoking cessation centers have been developed. During the 1-month pilot test, 20 volunteers participated and found no problems in transmitting text and multimedia messages. The participants were evaluated regarding the stage of behavioral change for each category via a simple questionnaire administered through the app (Multimedia Appendix 2) and messages tailored to the current stage were delivered.

Discussion

Rationale for Developing the AnSim App

The AnSim app was developed to support behavioral changes and decrease cardiovascular risk factors in patients who had undergone PCI. The messages of AnSim were developed based on the transtheoretical model of behavior change. A 5-phase systematic approach, from focus group interviews to the development of patient-specific message delivery systems, was conducted in collaboration with a multidisciplinary team.

Globally, CHD remains the major cause of death, despite advances in cardiovascular treatment. Further, the incidence rate of CHD is increasing in developing countries [3]. As an integral component of the continuum of cardiovascular care, secondary prevention and CR programs are recommended by most cardiovascular clinical guidelines as a Class I recommendation, and huge amounts of medical resources are devoted toward this endeavor. Although the clinical benefits [5,7] and cost-effectiveness [33] of CR programs have been reported, the supply and accessibility of CR programs are not satisfactory, especially in low-to-middle-income countries [34,35]. There are many hurdles that prevent patients from enrolling into CR programs, such as the distance from the patient's house to the CR center or a shortage of cost and time. Recently, the use of smartphones has increased worldwide, owing mainly to the development of mobile technology, and interest in mobile health care systems is increasing in various medical fields [14,19]. Smartphone-based CR can be a good alternative strategy that can enhance accessibility to medical care at a low cost [36].

Comparison With Prior Work

There had been many studies, although with varying number of participants, that showed the feasibility and positive results of mobile phone messaging in reducing body weight [37-39], increasing physical activity [40], and smoking cessation [41]. In particular, the Tobacco, Exercise and Diet Messages (TEXT ME) trial, one of the largest randomized controlled trials involving 710 patients with CHD, demonstrated that the use of an SMS text messaging service resulted in a modest improvement in the management of dyslipidemia and other cardiovascular disease risk factors [15]. These results are not surprising considering that home-based CR programs or education and counseling programs, which do not involve structured exercise therapy, show equivalent CHD prevention effects compared with traditional center-based CR or CR programs, including exercise programs [42].

With regard to smartphone apps, several small randomized studies, including patients with acute coronary syndrome having PCI, demonstrated improvement in treatment adherence [43] and weight loss [44] and a nonsignificant reduction in cardiovascular events [44]. The recent nonrandomized controlled trial with 1064 patients with acute myocardial infarction showed fewer all-cause 30 days readmissions in the digital intervention group compared with the control [45]. Unlike general concerns of smartphone-based interventions for the elderly, who account for a large proportion of patients with CHD, it has successfully improved physical activity and cognitive function in the older population [14,46]. Previous intervention strategies using mobile phones for CR were basically in a 1-way direction, and the contents were provided only as text messages and were not patient specific [8,15,27]. Through the AnSim app, the message is specific, tailored to the patient's behavioral stage after a brief review of recent medical records. The process is similar to that of a recent randomized controlled trial, the Smartphone and Social Media-Based Cardiac Rehabilitation and Secondary Prevention in China (SMART-CR/SP) trial, involving 312 patients with PCI [26].

Limitation

The 2-way direction system of the AnSim app is not complete, as it cannot directly answer or react to the patient's questions and needs immediately. However, the active interaction between CR apps and patients is expected to improve soon as artificial intelligence develops. Instead, the AnSim can deliver patient-specific messages that align with the step of each lifestyle category using the transtheoretical model of behavioral change and serial tracking of the status of patients. In addition, messages in the AnSim app can provide a variety of images, videos, sounds, and feedback, which can improve patient understanding and adherence and may allow for better effects.

Conclusions

In conclusion, this study reports the development of an app (AnSim) that provides a variety of medical information and CR programs regarding CHD. The messages were developed based on focus interviews, transtheoretical model, feedback, and refinement with various forms of multimedia, and the messages were intended to be specific to baseline characteristics and stage of behavioral change in each participant. Providing CR programs using mobile technology has a huge potential, and we expect that the AnSim app would be helpful for secondary prevention in patients who have undergone PCI. However, future studies are needed to determine the feasibility and efficacy of this app.

Acknowledgments

This research was supported by a grant from the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI16C0483). We thank Editage for English language editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The samples of message refinement.
[PNG File , 79 KB - medinform_v9i12e23285_app1.png]

Multimedia Appendix 2
Example of questionnaire for assessing behavioral change steps.
[DOCX File , 14 KB - medinform_v9i12e23285_app2.docx]

Multimedia Appendix 3
Baseline characteristics of patients participating in the focus group interview.
[DOCX File , 16 KB - medinform_v9i12e23285_app3.docx]

References

1. World Health Organization. Top 10 causes of death. Factsheet World Health Organization: World Health Organization. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [accessed 2021-07-05]
2. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012 Dec 15;380(9859):2095-2128. [doi: 10.1016/S0140-6736(12)61728-0] [Medline: 23245604]
3. Gaziano TA, Bitton A, Anand S, Abrahams-Gessel S, Murphy A. Growing epidemic of coronary heart disease in low- and middle-income countries. Curr Probl Cardiol 2010 Feb;35(2):72-115 [FREE Full text] [doi: 10.1016/j.cpcardiol.2009.10.002] [Medline: 20109979]
4. Briffa TG, Hobbs MS, Tonkin A, Sanfilippo FM, Hickling S, Ridout SC, et al. Population Trends of Recurrent Coronary Heart Disease Event Rates Remain High. Circ Cardiovasc Qual Outcomes 2011 Jan;4(1):107-113. [doi: 10.1161/circoutcomes.110.957944]
5. Dalal HM, Doherty P, Taylor RS. Cardiac rehabilitation. BMJ 2015 Sep 29;351:h5000 [FREE Full text] [Medline: 26419744]
6. Facts: cardiac rehabilitation: putting more patients on the road to recovery. American Heart Association. 2017 May. URL: https://www.heart.org/-/media/Files/About-Us/Policy-Research/Fact-Sheets/Clinical-and-Post-Clinical-Care/FACTS-Cardiac-Rehab.pdf [accessed 11/24/2021]
7. de Vries H, Kemps HM, van Engen-Verheul MM, Kraaijenhagen RA, Peek N. Cardiac rehabilitation and survival in a large representative community cohort of Dutch patients. Eur Heart J 2015 Apr 17;36(24):1519-1528. [doi: 10.1093/eurheartj/ehv111]
8. Persell SD, Peprah YA, Lipiszko D, Lee JY, Li JJ, Ciolino JD, et al. Effect of Home Blood Pressure Monitoring via a Smartphone Hypertension Coaching Application or Tracking Application on Adults With Uncontrolled Hypertension: A Randomized Clinical Trial. JAMA Netw Open 2020 Mar 02;3(3):e200255 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.0255] [Medline: 32119093]
9. Smith SC, Benjamin EJ, Bonow RO, Braun LT, Creager MA, Franklin BA, et al. AHA/ACCF Secondary Prevention and Risk Reduction Therapy for Patients With Coronary and Other Atherosclerotic Vascular Disease: 2011 Update. Circulation 2011 Nov 29;124(22):2458-2473. [doi: 10.1161/cir.0b013e318235eb4d]
10. Piepoli MF, Hoes AW, Agewall S, Albus C, Brotons C, Catapano AL, ESC Scientific Document Group. Eur Heart J 2016 Aug 01;37(29):2315-2381 [FREE Full text] [doi: 10.1093/eurheartj/ehw106] [Medline: 27222591]
11. Daly J, Sindone AP, Thompson DR, Hancock K, Chang E, Davidson P. Barriers to Participation in and Adherence to Cardiac Rehabilitation Programs: A Critical Literature Review. Progress in Cardiovascular Nursing 2007 Jun 15;17(1):8-17. [doi: 10.1111/j.0889-7204.2002.00614.x]
12. Dunlay SM, Witt BJ, Allison TG, Hayes SN, Weston SA, Koepsell E, et al. Barriers to participation in cardiac rehabilitation. American Heart Journal 2009 Nov;158(5):852-859. [doi: 10.1016/j.ahj.2009.08.010]
13. Redfern J, Ellis ER, Briffa T, Freedman SB. High risk - factor level and low risk - factor knowledge in patients not accessing cardiac rehabilitation after acute coronary syndrome. Medical Journal of Australia 2007 Jan;186(1):21-25. [doi: 10.5694/j.1326-5377.2007.tb00783.x]

14. Neubeck L, Lowres N, Benjamin EJ, Freedman SB, Coorey G, Redfern J. The mobile revolution--using smartphone apps to prevent cardiovascular disease. Nat Rev Cardiol 2015 Jun;12(6):350-360. [doi: 10.1038/nrcardio.2015.34] [Medline: 25801714]
15. Chow CK, Redfern J, Hillis GS, Thakkar J, Santo K, Hackett ML, et al. Effect of Lifestyle-Focused Text Messaging on Risk Factor Modification in Patients With Coronary Heart Disease: A Randomized Clinical Trial. JAMA 2015;314(12):1255-1263. [doi: 10.1001/jama.2015.10945] [Medline: 26393848]
16. Strandbygaard U, Thomsen SF, Backer V. A daily SMS reminder increases adherence to asthma treatment: a three-month follow-up study. Respir Med 2010 Feb;104(2):166-171 [FREE Full text] [doi: 10.1016/j.rmed.2009.10.003] [Medline: 19854632]
17. Adler AJ, Martin N, Mariani J, Tajer CD, Owolabi OO, Free C, et al. Mobile phone text messaging to improve medication adherence in secondary prevention of cardiovascular disease. Cochrane Database Syst Rev 2017 Apr 29;4:CD011851. [doi: 10.1002/14651858.CD011851.pub2] [Medline: 28455948]
18. Yudi MB, Clark DJ, Tsang D, Jelinek M, Kalten K, Joshi S, et al. SMARTphone-based, early cardiac REHABilitation in patients with acute coronary syndromes [SMART-REHAB Trial]: a randomized controlled trial protocol. BMC Cardiovasc Disord 2016 Dec 05;16(1):170 [FREE Full text] [doi: 10.1186/s12872-016-0356-6] [Medline: 27596569]
19. Varnfield M, Karunanithi M, Lee C, Honeyman E, Arnold D, Ding H, et al. Smartphone-based home care model improved use of cardiac rehabilitation in postmyocardial infarction patients: results from a randomised controlled trial. Heart 2014 Nov;100(22):1770-1779 [FREE Full text] [doi: 10.1136/heartjnl-2014-305783] [Medline: 24973083]
20. Li X, Yang S, Wang Y, Yang B, Zhang J. Effects of a transtheoretical model - based intervention and motivational interviewing on the management of depression in hospitalized patients with coronary heart disease: a randomized controlled trial. BMC Public Health 2020 Mar 30;20(1):420 [FREE Full text] [doi: 10.1186/s12889-020-08568-x] [Medline: 32228532]
21. Huang H, Lin Y, Chuang Y, Lin W, Kuo LY, Chen JC, et al. Application of the Transtheoretical Model to Exercise Behavior and Physical Activity in Patients after Open Heart Surgery. Acta Cardiol Sin 2015 May;31(3):202-208 [FREE Full text] [doi: 10.6515/acs20150204a] [Medline: 27122871]
22. Redfern J, Thiagalingam A, Jan S, Whittaker R, Hackett ML, Mooney J, et al. Development of a set of mobile phone text messages designed for prevention of recurrent cardiovascular events. Eur J Prev Cardiol 2014 Apr;21(4):492-499. [doi: 10.1177/2047487312449416] [Medline: 22605787]
23. Zhang H, Jiang Y, Nguyen HD, Poo DCC, Wang W. The effect of a smartphone-based coronary heart disease prevention (SBCHDP) programme on awareness and knowledge of CHD, stress, and cardiac-related lifestyle behaviours among the working population in Singapore: a pilot randomised controlled trial. Health Qual Life Outcomes 2017 Mar 14;15(1):49 [FREE Full text] [doi: 10.1186/s12955-017-0623-y] [Medline: 28288636]
24. Lear SA, Ignaszewski A. Cardiac rehabilitation: a comprehensive review. Curr Control Trials Cardiovasc Med 2001;2(5):221-232 [FREE Full text] [doi: 10.1186/cvm-2-5-221] [Medline: 11806801]
25. Kim C, Sung J, Lee JH, Kim WS, Lee GJ, Jee S, et al. Clinical Practice Guideline for Cardiac Rehabilitation in Korea: Recommendations for Cardiac Rehabilitation and Secondary Prevention after Acute Coronary Syndrome. Korean Circ J 2019 Nov;49(11):1066-1111 [FREE Full text] [doi: 10.4070/kcj.2019.0194] [Medline: 31646772]
26. Dorje T, Zhao G, Tso K, Wang J, Chen Y, Tsokey L, et al. Smartphone and social media-based cardiac rehabilitation and secondary prevention in China (SMART-CR/SP): a parallel-group, single-blind, randomised controlled trial. The Lancet Digital Health 2019 Nov;1(7):e363-e374. [doi: 10.1016/s2589-7500(19)30151-7]
27. Zheng X, Spatz ES, Bai X, Huo X, Ding Q, Horak P, et al. Effect of Text Messaging on Risk Factor Management in Patients With Coronary Heart Disease: The CHAT Randomized Clinical Trial. Circ Cardiovasc Qual Outcomes 2019 Apr;12(4):e005616. [doi: 10.1161/CIRCOUTCOMES.119.005616] [Medline: 30998400]
28. Green HE. Use of theoretical and conceptual frameworks in qualitative research. Nurse Res 2014 Jul;21(6):34-38. [doi: 10.7748/nr.21.6.34.e1252] [Medline: 25059086]
29. Varpio L, Paradis E, Uijtdehaage S, Young M. The Distinctions Between Theory, Theoretical Framework, and Conceptual Framework. Acad Med 2020 Jul;95(7):989-994. [doi: 10.1097/ACM.0000000000003075] [Medline: 31725464]
30. Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. Health Psychol 2008 May;27(3):379-387. [doi: 10.1037/0278-6133.27.3.379] [Medline: 18624603]
31. Skinner B. About Behaviorism. Pimlico, UK: Vintage; 2011:0307797848.
32. Prochaska JO, Velicer WF. The transtheoretical model of health behavior change. Am J Health Promot 1997 Aug 26;12(1):38-48. [doi: 10.4278/0890-1171-12.1.38] [Medline: 10170434]
33. Wong WP, Feng J, Pwee KH, Lim J. A systematic review of economic evaluations of cardiac rehabilitation. BMC Health Serv Res 2012 Aug 8;12(1):1-8. [doi: 10.1186/1472-6963-12-243]
34. Turk-Adawi K, Sarrafzadegan N, Grace SL. Global availability of cardiac rehabilitation. Nat Rev Cardiol 2014 Jul 15;11(10):586-596. [doi: 10.1038/nrcardio.2014.98]
35. Shanmugasegaram S, Perez-Terzic C, Jiang X, Grace SL. Cardiac rehabilitation services in low- and middle-income countries: a scoping review. J Cardiovasc Nurs 2014;29(5):454-463. [doi: 10.1097/JCN.0b013e31829c1414] [Medline: 23839574]

36. Commission E. Green paper on mobile health ("mHealth"). European Commission. 2014. URL: https://digital-strategy.ec.europa.eu/en/library/green-paper-mobile-health-mhealth#:~:text=Green%20Paper%20on%20mobile%20health%20%28%22mHealth %22%29%20On%2010,to%20the%20uptake%20of%20mHealth%20in%20the%20EU [accessed 2014-04-10]
37. Patrick K, Raab F, Adams MA, Dillon L, Zabinski M, Rock CL, et al. A text message-based intervention for weight loss: randomized controlled trial. J Med Internet Res 2009;11(1):e1 [FREE Full text] [doi: 10.2196/jmir.1100] [Medline: 19141433]
38. Gerber BS, Stolley MR, Thompson AL, Sharp LK, Fitzgibbon ML. Mobile phone text messaging to promote healthy behaviors and weight loss maintenance: a feasibility study. Health Informatics J 2009 Mar;15(1):17-25 [FREE Full text] [doi: 10.1177/1460458208099865] [Medline: 19218309]
39. Joo N, Kim B. Mobile phone short message service messaging for behaviour modification in a community-based weight control programme in Korea. J Telemed Telecare 2007;13(8):416-420. [doi: 10.1258/135763307783064331] [Medline: 18078554]
40. Shapiro JR, Bauer S, Hamer RM, Kordy H, Ward D, Bulik CM. Use of text messaging for monitoring sugar-sweetened beverages, physical activity, and screen time in children: a pilot study. J Nutr Educ Behav 2008;40(6):385-391 [FREE Full text] [doi: 10.1016/j.jneb.2007.09.014] [Medline: 18984496]
41. Liao Y, Wu Q, Kelly BC, Zhang F, Tang Y, Wang Q, et al. Effectiveness of a text-messaging-based smoking cessation intervention ("Happy Quit") for smoking cessation in China: A randomized controlled trial. PLoS Med 2018 Dec;15(12):e1002713 [FREE Full text] [doi: 10.1371/journal.pmed.1002713] [Medline: 30562352]
42. Clark AM, Hartling L, Vandermeer B, McAlister FA. Meta-Analysis: Secondary Prevention Programs for Patients with Coronary Artery Disease. Ann Intern Med 2005 Nov 01;143(9):659. [doi: 10.7326/0003-4819-143-9-200511010-00010]
43. Johnston N, Bodegard J, Jerström S, Åkesson J, Brorsson H, Alfredsson J, et al. Effects of interactive patient smartphone support app on drug adherence and lifestyle changes in myocardial infarction patients: A randomized study. Am Heart J 2016 Aug;178:85-94 [FREE Full text] [doi: 10.1016/j.ahj.2016.05.005] [Medline: 27502855]
44. Widmer RJ, Allison TG, Lennon R, Lopez-Jimenez F, Lerman LO, Lerman A. Digital health intervention during cardiac rehabilitation: A randomized controlled trial. Am Heart J 2017 Jun;188:65-72. [doi: 10.1016/j.ahj.2017.02.016] [Medline: 28577682]
45. Marvel FA, Spaulding EM, Lee MA, Yang WE, Demo R, Ding J, et al. Digital Health Intervention in Acute Myocardial Infarction. Circ Cardiovasc Qual Outcomes 2021 Jul;14(7):e007741 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.121.007741] [Medline: 34261332]
46. Zhang L, Zhang L, Wang J, Ding F, Zhang S. Community health service center-based cardiac rehabilitation in patients with coronary heart disease: a prospective study. BMC Health Serv Res 2017 Feb 11;17(1):128 [FREE Full text] [doi: 10.1186/s12913-017-2036-3] [Medline: 28187728]

Abbreviations

CHD: coronary heart disease
CR: cardiac rehabilitation
PCI: percutaneous coronary intervention
PCSK9: proprotein convertase subtilisin/kexin type 9

Edited by C Lovis; submitted 08.08.20; peer-reviewed by R Krukowski, J Li, MDG Pimentel, N Mohammad Gholi Mezerji; comments to author 07.10.20; revised version received 27.03.21; accepted 10.10.21; published 07.12.21.

Please cite as:
Choi JY, Kim JB, Lee S, Lee SJ, Shin SE, Park SH, Park EJ, Kim W, Na JO, Choi CU, Rha SW, Park CG, Seo HS, Ahn J, Jeong HG, Kim EJ
A Smartphone App (AnSim) With Various Types and Forms of Messages Using the Transtheoretical Model for Cardiac Rehabilitation in Patients With Coronary Artery Disease: Development and Usability Study
JMIR Med Inform 2021;9(12):e23285
URL: https://medinform.jmir.org/2021/12/e23285
doi:10.2196/23285
PMID:34878987

Original Paper

The Effect of Automated Mammogram Orders Paired With Electronic Invitations to Self-schedule on Mammogram Scheduling Outcomes: Observational Cohort Comparison

Frederick North[1], MD; Elissa M Nelson[2], MA; Rebecca J Buss[2], MS; Rebecca J Majerus[2], BAS; Matthew C Thompson[2], MBA; Brian A Crum[3], MD

[1]Division of Community Internal Medicine, Department of Internal Medicine, Mayo Clinic, Rochester, MN, United States

[2]Enterprise Office of Access Management, Mayo Clinic, Rochester, MN, United States

[3]Department of Neurology, Mayo Clinic, Rochester, MN, United States

Corresponding Author:
Frederick North, MD
Division of Community Internal Medicine
Department of Internal Medicine
Mayo Clinic
200 First Street SW
Rochester, MN, 55905
United States
Phone: 1 507 284 2511
Email: north.frederick@mayo.edu

Abstract

Background: Screening mammography is recommended for the early detection of breast cancer. The processes for ordering screening mammography often rely on a health care provider order and a scheduler to arrange the time and location of breast imaging. Self-scheduling after automated ordering of screening mammograms may offer a more efficient and convenient way to schedule screening mammograms.

Objective: The aim of this study was to determine the use, outcomes, and efficiency of an automated mammogram ordering and invitation process paired with self-scheduling.

Methods: We examined appointment data from 12 months of scheduled mammogram appointments, starting in September 2019 when a web and mobile app self-scheduling process for screening mammograms was made available for the Mayo Clinic primary care practice. Patients registered to the Mayo Clinic Patient Online Services could view the schedules and book their mammogram appointment via the web or a mobile app. Self-scheduling required no telephone calls or staff appointment schedulers. We examined uptake (count and percentage of patients utilizing self-scheduling), number of appointment actions taken by self-schedulers and by those using staff schedulers, no-show outcomes, scheduling efficiency, and weekend and after-hours use of self-scheduling.

Results: For patients who were registered to patient online services and had screening mammogram appointment activity, 15.3% (14,387/93,901) used the web or mobile app to do either some mammogram self-scheduling or self-cancelling appointment actions. Approximately 24.4% (3285/13,454) of self-scheduling occurred after normal business hours/on weekends. Approximately 9.3% (8736/93,901) of the patients used self-scheduling/cancelling exclusively. For self-scheduled mammograms, there were 5.7% (536/9433) no-shows compared to 4.6% (3590/77,531) no-shows in staff-scheduled mammograms (unadjusted odds ratio 1.24, 95% CI 1.13-1.36; P<.001). The odds ratio of no-shows for self-scheduled mammograms to staff-scheduled mammograms decreased to 1.12 (95% CI 1.02-1.23; P=.02) when adjusted for age, race, and ethnicity. On average, since there were only 0.197 staff-scheduler actions for each finalized self-scheduled appointment, staff schedulers were rarely used to redo or “clean up” self-scheduled appointments. Exclusively self-scheduled appointments were significantly more efficient than staff-scheduled appointments. Self-schedulers experienced a single appointment step process (one and done) for 93.5% (7553/8079) of their finalized appointments; only 74.5% (52,804/70,839) of staff-scheduled finalized appointments had a similar one-step appointment process (P<.001). For staff-scheduled appointments, 25.5% (18,035/70,839) of the finalized appointments took multiple appointment steps. For finalized appointments that were exclusively self-scheduled, only 6.5% (526/8079) took multiple appointment steps.

The staff-scheduled to self-scheduled odds ratio of taking multiple steps for a finalized screening mammogram appointment was 4.9 (95% CI 4.48-5.37; *P*<.001).

Conclusions: Screening mammograms can be efficiently self-scheduled but may be associated with a slight increase in no-shows. Self-scheduling can decrease staff scheduler work and can be convenient for patients who want to manage their appointment scheduling activity after business hours or on weekends.

(JMIR Med Inform 2021;9(12):e27072) doi:10.2196/27072

KEYWORDS

electronic health record; schedule; patient appointment; preventive health service; office visit; outpatient care; mammogram; software tool; computer software application; mobile applications; self-schedule; app; EHR; screening; diagnostic; cancer

Introduction

About 1 in 8 women in the United States will develop breast cancer during her life [1]. Breast cancer screening with mammograms can help detect breast cancer at an early stage when treatment is most successful [2]. Despite the need for breast cancer screening, 31% of women in the screening age range of 45-55 years have not had a mammogram in the last 2 years [3]. Several interventions have been tried to increase the percentage of women receiving screening mammograms [4]. Primary care health care providers have historically played a major role in advising patients about screening mammography. Typically, providers address preventive health services, including screening mammography, during the periodic examination [5]. However, despite some continued promotion of the periodic health care examination [6], there is no overwhelming evidence for the periodic examination to significantly change health outcomes, including breast cancer [7]. In addition, screening mammography is often just one of many recommended actions that primary care providers need to address with their patients. In a study at Mayo Clinic, we found that primary care patients aged 50-65 years, on average, had 5.5 unmet health care recommendations, with the conclusion that there needs to be “new approaches to address the burgeoning numbers of uncompleted recommendations” [8]. Yarnall et al [9] also noted the large amount of time that is required for primary care providers to address every preventive service, including screening mammography.

Automated ordering and self-scheduling of screening mammography with the assistance of the electronic health record (EHR) is an intervention that could help deliver the preventive service of early breast cancer detection in a primary care practice. Criteria for screening mammography can often be found within the EHR. For example, the American Cancer Society recommends mammograms up to every year for women aged 40-75 years depending on the life expectancy [2]. Determining whether a screening mammogram is due for a given individual can be accomplished through software rules that query the EHR for patient characteristics and dates of previous mammograms. Self-scheduling has been used for airline, hotel, and event bookings for years. So why has the self-scheduling of medical appointments lagged? The short answer is that medical appointments encompass many different appointment types and appointment purposes that require very different rules for scheduling. For example, Zocdoc.com is an internet third party medical appointment enabler that matches individuals on the web with health care providers for scheduled visits. Zocdoc makes some of the details of the matching process available [10,11]. Scheduling in Zocdoc includes very specific rules such as matching insurance coverage, preferred medical specialty, and availability for face-to-face or video visit [12,13]. COVID-19 visits are another very specific visit type requiring specific criteria for booking. In a recent study, Judson et al [14] noted how self-triage rules in the self-scheduling process were designed to limit COVID video visits to those who did not require more emergent care. Because of the differences in appointment purpose and type, the COVID-19 self-scheduling rules are very different from those used by Zocdoc for more general appointments. The periodic well-child examination is another example of a self-scheduling appointment type that requires a completely different set of rules. Scheduling of the well-child examination is based on the age of the child, the date of the last well-child examination, and matching with the child’s primary care provider [15].

The screening mammogram appointment is also a visit type with its own unique set of rules that distinguish it from other visit types. The unique challenges for self-scheduling screening mammograms are (1) there are specific criteria for patient age, date of the last mammogram, and whether a screening mammogram is appropriate; (2) it is a radiologic procedure requiring an electronic order; (3) there are patient and provider requirements so that the assignment and communication of results is assured. In addition to examining the outcomes of self-scheduled mammograms, we show our automated processes for the self-scheduled screening mammogram visit that address the unique challenges of this visit type.

Methods

Setting

The implementation of automated ordering paired with self-scheduling of mammograms took place at Mayo Clinic in 2019. Mayo Clinic is a multispecialty group practice with several locations in the United States and internationally. Our study focuses on the screening mammogram process of the primary care practices of Mayo Clinic for 12 consecutive months from September 2019 through August 2020. Mayo Clinic has primary care practices in the United States in Florida, Arizona, and many locations in the upper Midwest, primarily in the states of Minnesota, Wisconsin, and Iowa. All the primary care sites were included in this study. This study was limited to the bilateral breast screening mammogram examination, which is

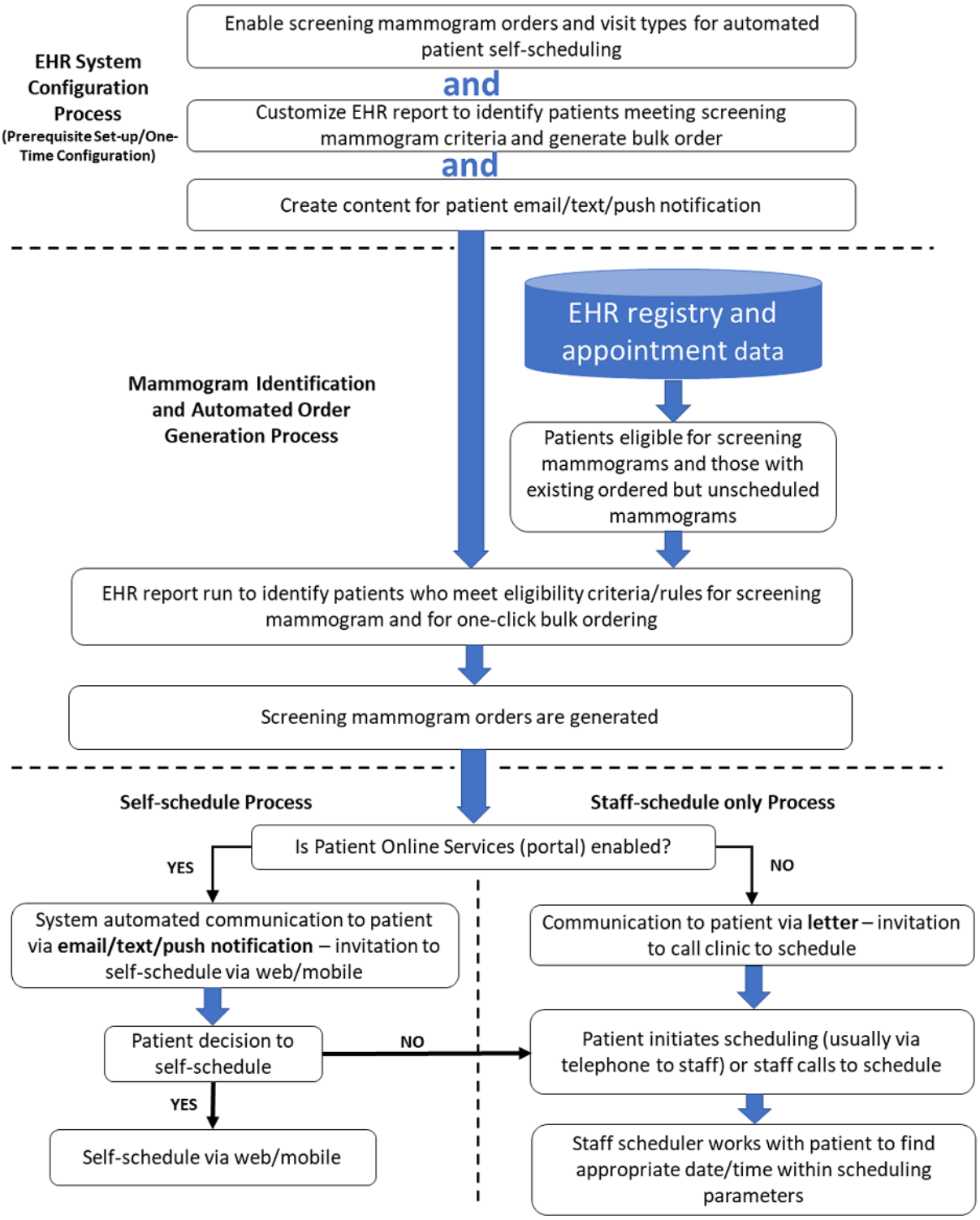
the recommended radiographic procedure for the early detection of breast cancer.

Automated Screening Mammogram Order Process

At Mayo Clinic, screening mammography requires an order for the specific imaging examination. Patients are not allowed to self-order a mammogram. However, Mayo Clinic has developed a process with rules that automatically generates orders for screening mammograms, allowing one-click bulk ordering of hundreds of mammograms by a single provider. The top part of Figure 1 shows the one-time EHR system configuration set-up needed for the mammogram bulk ordering process. The configuration of the order and visit types in the scheduling system were needed to allow automated mammogram ordering. A special EHR report was configured to identify patients meeting the screening mammogram criteria and to produce the bulk mammogram order. Creation of patient email/text/push notification content was also required for the self-schedule electronic invitation process.

After the prerequisite EHR system configurations are completed, the mammogram ordering process starts by using EHR data to identify patients who are eligible and due for screening mammography. The appointment scheduling system is also queried for those who are due and have a mammogram ordered but not scheduled. For those who do not have an active mammogram order but are due for a mammogram, a mammogram order is created. Thus, all patients who are due for a mammogram either have a mammogram order generated automatically to enable scheduling or they are identified to enable scheduling if an active mammogram order is already in place but not yet scheduled. Figure 1 (bottom third) shows that once the mammogram order is generated or identified as needing scheduling, the process diverges depending on whether the patient is enabled with patient online services. All patients who are due and had the mammogram order generated or who have an existing mammogram order needing scheduling are sent invitations to schedule their mammograms. Those who use patient online services are sent invitations by an email message and, if mobile app, a push notification. Those without patient online services are sent a letter by post. The mammogram invitations sent by the portal included an invitation to self-schedule. All those with patient online services are enabled to self-schedule both by using web and mobile app. Patients with patient online services also have the option to have staff help them schedule their mammograms (staff scheduled) via a phone call or portal message. For patients without patient online services, mammograms can only be scheduled with staff assistance (staff scheduler).

Figure 1. Prerequisite system configuration and process flow for automated identification of eligible patients for screening mammograms, automated mammogram order generation, and communication to patients for self-scheduling versus staff scheduling. EHR: electronic health record.



Staff Scheduling Versus Self-scheduling

Staff schedulers are clinic staff employees who schedule or cancel appointments for patients. Until the self-scheduling process was implemented, staff employee appointment schedulers were responsible for working with patients and radiology schedules to schedule mammograms. Patients, whether patient online services–enabled or not, can schedule their mammograms by telephone or in person via staff appointment schedulers. Appointment scheduling via staff schedulers normally occurs during business hours of 7 AM to 5 PM on weekdays. Appointment schedulers have the ability to schedule mammograms more than 12 weeks into the future. Self-scheduling via patient online services can be done either via web or via mobile and is available 24/7. Patients can directly see the mammogram scheduling template for the days that they

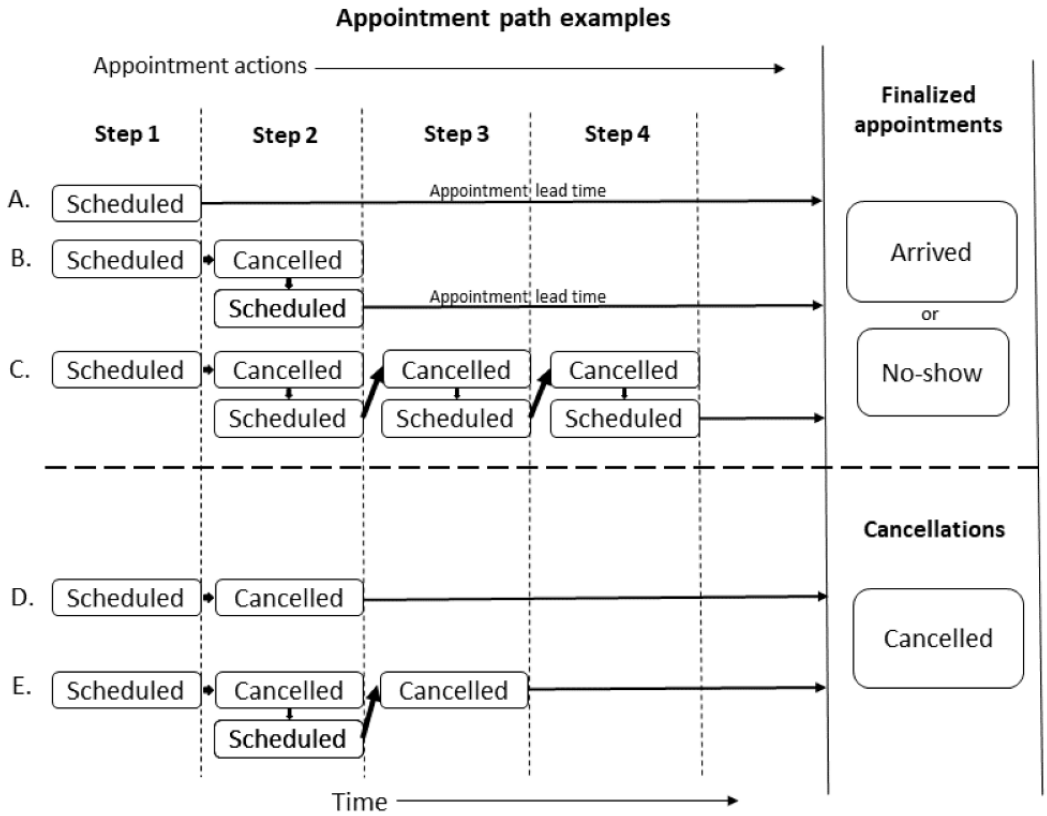
select and can click on the appointment time that they want. Self-schedulers are restricted to scheduling their mammogram in a 12-week rolling window from the day that they could schedule. Self-schedulers are also not allowed to double book.

Appointment Definitions

Self-schedulers or self-cancelers are the patients who used the Mayo software interface (web or mobile) to self-schedule or self-cancel the mammogram appointments. It should be noted that we focus on self-schedule actions in this study. There were some patients who never used the self-scheduling feature but self-cancelled the appointments made by the staff schedulers. To be considered self-scheduled, a patient had to have at least one appointment action of self-scheduling (booking an appointment with the self-schedule software). The few patients who self-cancelled their staff-scheduled appointment were classified as staff-scheduled.

An appointment action is either a schedule or cancel event. With the self-scheduling process, appointment actions could be done either by the patient (self) or staff. An appointment path is the sequence of appointment actions leading to a finalized appointment or cancellation outcome (Figure 2). Appointment paths can contain both self and staff appointment actions. The example above of a self-cancelled appointment that was scheduled by a staff would have 2 appointment actions: a staff-scheduler action and a self-cancel action.

Figure 2. Examples of different appointment paths showing the appointment actions and appointment steps leading to a finalized appointment or cancellation.



Finalized appointments were those scheduled appointments that were left scheduled up to the appointment date and time (not cancelled before appointment time). Figure 2 shows examples of appointment paths and appointment outcomes. Our data start with a time-stamped appointment schedule action. We dichotomized appointment actions into those by staff schedulers and those by self-schedulers. As shown in Figure 2, each patient (whether self-scheduled or staff-scheduled) begins with a scheduling action that we term as appointment step 1. Patients can then go through several decision steps of whether to cancel or reschedule (a cancel and schedule pair). Some patients would reschedule multiple times before a finalized appointment. To quantify this activity, we counted the appointment steps. Figure 2A shows an appointment path to appointment finalization with just 1 step, the initial scheduling action. Figure 2B and Figure 2C show appointment paths to appointment finalization taking 2 and 4 steps, respectively. Appointment paths ending in a cancellation outcome also may take several appointment steps. Figure 2D and Figure 2E show cancellation examples that take 2 and 3 appointment steps, respectively, to result in a cancellation.

Appointment outcomes are dichotomously categorized as finalized appointments or cancellations. Finalized appointments are further dichotomously categorized as completed or no-shows (never arrived at the scheduled appointment time). Figure 2A also shows the appointment lead time, which is the scheduled appointment date/time minus the date/time the appointment was made. This is the lead time that the patient has from the date of scheduling the appointment to the actual future-reserved appointment date.

Mammogram Appointment Selection and Follow-up

Our data source was EHR-generated scheduling and cancelling information on all bilateral screening mammogram appointments made for primary care patients for the 12 months from September 1, 2019 through August 31, 2020. Scheduled mammogram appointments were either cancelled or completed (patient arrived or no-show) by September 1, 2020 for our follow-up on finalized appointments and no-shows. Patients eligible for automated ordering and invitation to schedule a mammogram were primary care practice patients of Mayo Clinic in Arizona or Florida or in the Mayo Clinic Health System (Minnesota, Wisconsin, Iowa). Self-scheduling through web or mobile required patient online service registration; staff scheduling was available for all patients who had an active mammogram order, regardless of patient online service registration status.

Summary of the Outcome Measures

A finalized scheduled mammogram was the outcome of scheduling and cancelling actions as shown in Figure 2. A finalized mammogram appointment was defined as a mammogram appointment scheduled and remaining active until the date and time of the scheduled mammogram radiology visit. Scheduling and cancelling actions were outcomes of interest defined as the scheduling of a mammogram (booking an assigned time and date for the mammogram) or the action of cancelling a mammogram (cancelling a previously booked mammogram appointment). Scheduling and cancelling actions were dichotomized depending on whether they were accomplished by self-scheduling or by staff schedulers. The no-show mammogram appointment, defined as the finalized appointment for a patient who never arrived for their scheduled mammogram, was also an outcome of interest. Appointment lead time was defined as the time difference between the actual appointment date and time and the date and time it was last scheduled, after any prior schedule and cancel actions as noted in Figure 2. Appointment lead times were of interest because staff schedulers could schedule mammogram appointments beyond the 12-week lead time limit of self-scheduling. Patient uptake of the self-scheduling process was measured as counts and percentage of patients over time who used self-scheduling or a combination of self-scheduling and self-cancelling exclusively or in combination with staff scheduling for their appointment actions for a finalized appointment. The mutually exclusive 3 categories of patients who finalized appointments were as follows: self-scheduled exclusively (could also self-cancel), self- and staff-scheduled (any combination of self and staff scheduling actions), and patients who used staff schedulers exclusively (no self-scheduling or self-cancelling appointment actions).

Data Collection and Analysis

Data were collected by the Epic EHR of Mayo Clinic. Patients who either staff-scheduled or self-scheduled were registered patients of Mayo Clinic. In addition, we limited this study to portal-registered patients who were established primary care patients; therefore, there were essentially complete demographic data available for each patient (age, race, sex, ethnicity). Any uncategorized or missing information on race or ethnicity was placed in the other or unknown category. Appointment data were entered by the Epic scheduling software. Dates and times of self-scheduling and staff scheduling were automatically entered into the EHR software by patient record number and categorized on data entry as being sourced from self-scheduling (patient online services) or by the staff scheduler. Mammograms were not done unless the patient was checked in by radiology staff as “arrived.” If the patient did not arrive and the radiology staff overlooked listing the patient as a no-show, an EHR scheduling rule marked the visit as a no-show 72 hours after the scheduled appointment to ensure the capture of these overlooked no-shows.

We categorized scheduling and cancelling actions according to whether they occurred outside of the usual business hours (Monday through Friday, 7 AM to 5 PM). The proportion of mammogram appointment lead times over 12 weeks was calculated for both staff-scheduled and self-scheduled appointments. As mentioned above, only those scheduled for mammography who were registered with patient online services were analyzed. Thus, portal registration status was not in our primary analysis. Those without portal registration were included in additional analysis as described below. Age is a known confounder for no-shows in radiology visits [16]; therefore, we adjusted for age in our analysis of no-shows. We conducted additional analyses to determine how sensitive our findings were to the disruption that the COVID-19 pandemic had on mammogram appointments. In March 2020, shortly after the midpoint of our data capture, mammogram appointments were suspended temporarily. It was unclear how much this disruption of scheduling affected self-scheduling activity and whether it increased the use of staff schedulers. To quantify this, we analyzed separately the 6 pre-COVID months and the 6 post-COVID months (September 2019 through February 2020 and March 2020 through August 2020, respectively) for self-scheduling and staff-scheduling activity. We also performed additional data analysis to evaluate the self-scheduling uptake for all patients scheduling mammograms, including those without portal registration.

Statistical Analysis

We used JMP 14.3 (SAS Institute Inc) for the statistical analysis. The chi-square test was used for categorical analysis. We used logistic regression analysis in a model to explain the differences in the no-shows adjusted by patient age to control for age as a known confounder in radiology no-shows [16]. A logistic regression analysis model using age, race, and ethnicity was also used to adjust for additional differences in demographics for the no-shows analysis.

Ethics

This was a retrospective study examining quality measures and uptake of a self-scheduling process. Self-scheduling was a voluntary additional option offered to all primary care patients with patient online services; all individuals could continue to schedule their mammograms with staff schedulers if that was their preference (see patient decision point in Figure 1). This study met the institutional review board criteria for exemption (IRB-2020-006809).

Results

Uptake of Self-scheduling Mammograms

Figure 3 shows the patient counts of those who had mammograms scheduled for the 12 months of the study. Approximately 16.4% (18,466/112,367) of the patients did not have access to either self-scheduling or self-cancelling (not registered with patient online services). In this study, we focused on 93,901 individuals who had access to self-scheduling. Of those individuals, 15.3% (14,387/93,901) used self-scheduling or self-cancelling. Of those with patient online services, 9.3% (8736/93,901) exclusively used self-scheduling and thus did not use any staff-scheduler resources. Another 6% (5651/93,901) used some self-scheduling/self-cancelling processes to arrange their screening mammogram.

Figure 3. Patients who had scheduling actions for bilateral screening mammograms for the 12 months of the study. Patient counts show those who exclusively used self-scheduling, those exclusively staff-scheduled, and those who had both self-scheduling and staff-scheduling appointment actions.

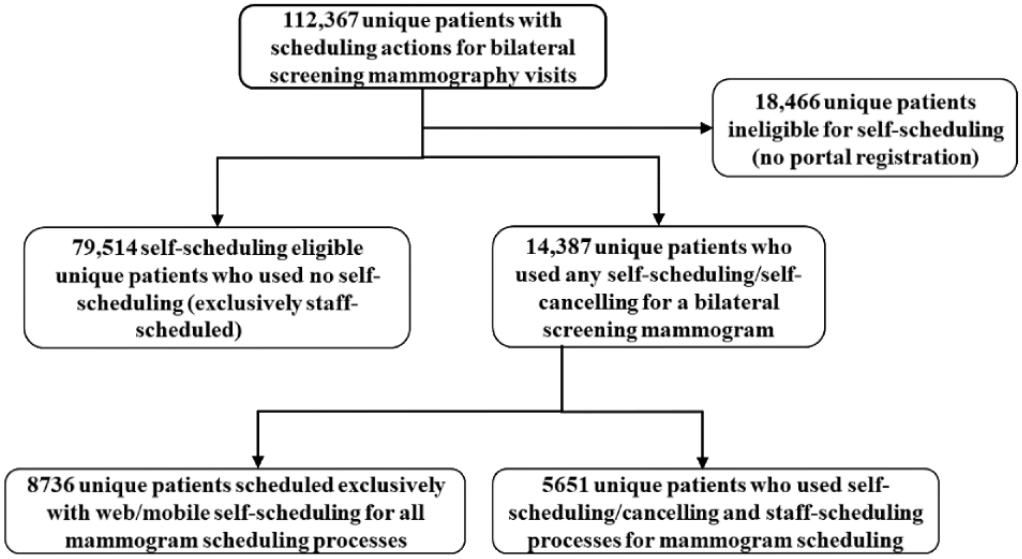
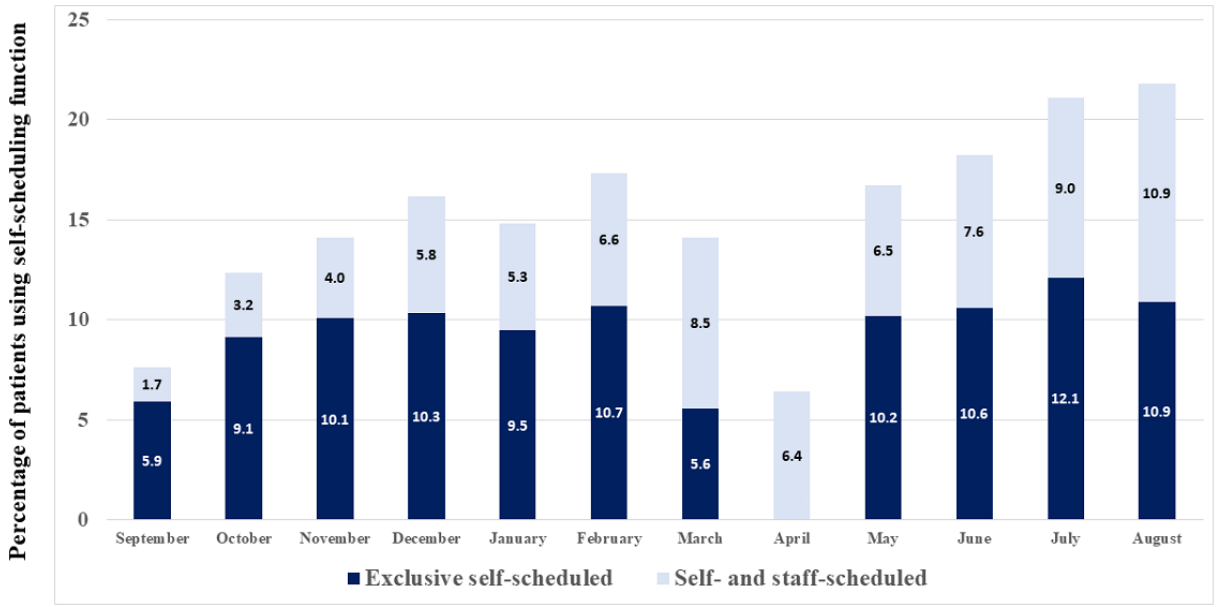


Figure 4 shows the longitudinal percentage uptake of self-scheduling for those who had self-scheduling access. In the initial month of widespread implementation, 7.6% (678/8898) of all individuals involved in scheduling mammograms were doing some self-scheduling actions. Eleven months later (July 2020), this had increased by 276%, so that 21.1% (1991/9442) of the patients scheduling mammograms were doing some self-scheduling. At 12 months (August 2020), 21.8% (1091/5005) of the patients were doing some self-scheduling, but since many patients who started scheduling in August had not reached the scheduled date of their appointment (finalized their appointment) by the end of data collection (August 31, 2020), the counts were lower. The drop in scheduling in March and April 2020 was associated with access limitations imposed during the initial months of the COVID-19 pandemic. As part of those restrictions, self-scheduling was not available for scheduling mammograms during part of March 2020 and all of April 2020, but self-cancelling was still available.

Figure 4. Longitudinal uptake of self-scheduling paired with automatically generated invitations to schedule mammograms (September 2019 to August 2020). The graph shows the percentage of patients with patient online services–enabled who either exclusively used self-scheduling or used some self-scheduling. Self-cancelling activity took place in April 2020 when patients could not self-schedule.



Demographics of the Patients

Table 1 compares the demographics of the individuals who had patient online services and performed any self-scheduling activity with those of individuals who had staff-scheduled appointments. There were notable differences in the age distributions, which was consistent with younger individuals being more comfortable with web and mobile technology. Although statistically there were some racial differences, the absolute percentages were similar. The percentage of White females in the self-scheduled was 93.7% (13,474/14,382) compared to 93.7% (74,436/79,476) in the staff scheduled, thereby showing a nonsignificant difference (P=.90).

Table 1. Demographics of individuals who used self-scheduling compared to those of individuals who used staff-scheduling for making appointments for their screening mammograms.

Demographic characteristic	Any self-scheduled, (n=14,387), n (%)	Exclusively staff-scheduled (n=79,514), n (%)	*P* value[a]
Age (years)			<.001
20-29	2 (0.01)	40 (0.05)	
30-39	91 (0.63)	606 (0.76)	
40-49	4311 (29.96)	15,113 (19.01)	
50-59	4468 (31.06)	21,322 (26.82)	
60-69	3954 (27.48)	24,977 (31.41)	
70-79	1408 (9.79)	14,675 (18.46)	
80-89	148 (1.03)	2674 (3.36)	
90-99	5 (0.03)	107 (0.13)	
Self-described gender (female)	14,382 (99.97)	79,476 (99.95)	.50
Race			.002
White	13,474 (93.65)	74,436 (93.61)	
Black	186 (1.29)	1357 (1.71)	
Asian	316 (2.20)	1577 (1.98)	
Other	269 (1.87)	1420 (1.79)	
Not disclosed	142 (0.99)	724 (0.91)	
Ethnicity			<.001
Hispanic	336 (2.34)	2339 (2.94)	
Not Hispanic	13,772 (95.73)	75,745 (95.26)	
Undisclosed/unknown	279 (1.94)	1430 (1.80)	

[a]Null hypothesis (H0) tested: percentage of each demographic characteristic is equal between those who performed any self-scheduled activity and those who had staff-scheduled appointments exclusively.

Appointment Actions Completed by Self-schedulers

As mentioned in Methods, before an appointment is finalized, it can be cancelled and rescheduled many times. Table 2 shows the counts of all the scheduling and cancelling appointment actions done by self-scheduled patients and those done by staff schedulers. Out of 175,256 appointment actions completed, 10% (17,475/175,256) were done by patients. All the appointment actions resulted in a total of 86,964 finalized appointments, with 10.8% (9433/86,964) at least partially finalized by the patient.

Table 2. Appointment metric comparison between self-scheduled and staff-scheduled appointments for those with access to self-scheduling (patient online services–enabled).

Appointment metric	Self-scheduled but staff could cancel	Staff-scheduled but patients could still self-cancel	*P* value[a]
Appointment actions, n (%)			
Self-scheduled	13,454 (100)	0 (0)	<.001
Staff-scheduled	0 (0)	117,656 (100)	<.001
Self-cancelled	2166 (16.10)	3847 (3.27)	<.001
Staff-cancelled	1855 (13.79)	36,278 (30.83)	<.001
Total cancelled	4021 (29.89)	40,125 (34.10)	<.001
Appointment outcomes, n (%)			
Finalized appointments (scheduled minus cancelled)	9433 (100)	77,531 (100)	N/A[b]
Arrived to appointment	8897 (94.32)	73,941 (95.37)	<.001
No-show	536 (5.68)	3590 (4.63)	<.001
Appointment action efficiency			
Total appointment actions per finalized appointment (total count of the above 4 rows of self-scheduling and staff-scheduling and cancelling appointment actions divided by the total count of finalized appointments)	1.852	2.035	N/A
Self-generated appointment actions per finalized appointment (total count of the above 2 rows of self-scheduled and self-cancelled appointment actions divided by the total count of finalized appointments)	1.656	0.050	N/A
Staff-generated appointment actions per finalized appointment (total count of the above 2 rows of Mayo staff-scheduled and staff-cancelled appointment actions divided by the total count of finalized appointments)	0.197	1.985	N/A
Appointment actions outside of standard appointment scheduler hours, n (%)			
Scheduling actions completed outside of normal business hours of Monday to Friday, 7 AM to 5 PM	3285 (24.42)	1659 (1.41)	<.001
Scheduling actions completed on Saturday or Sunday	1149 (8.54)	769 (0.65)	<.001
Scheduling actions completed on Monday to Friday outside of 7 AM to 5 PM	2136 (15.88)	890 (0.76)	<.001
Appointment lead time			
Median lead time (days)	15	21	N/A
Lead time over 84 days, n (%)	0 (0)	5778 (4.91)	<.001

[a]Null hypothesis (H0) tested: proportion of self-scheduled appointments equals staff-scheduled appointments.

[b]N/A: not applicable.

Convenience of Scheduling

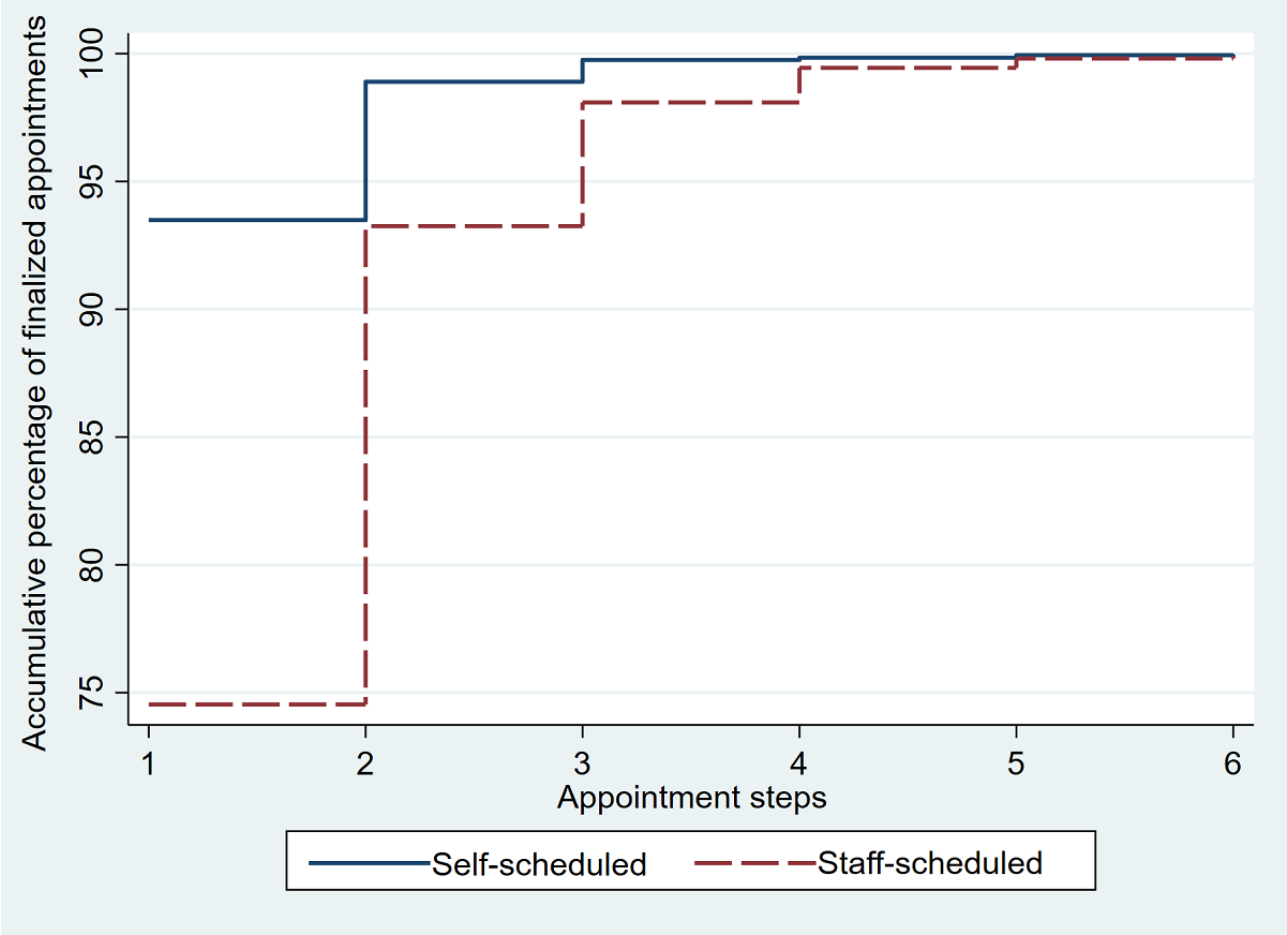
Approximately 24.4% (3285/13,454) of the mammogram self-scheduling activity was accomplished either on the weekend or on weekdays after usual staff scheduler hours (Table 2). This after-hours scheduling was done during the weekday for 15.9% (2136/13,454) of the appointment actions and on the weekend for 8.5% (1149/13,454) of the appointment actions. Approximately 75.5% (10,163/13,454) of the self-scheduling appointment actions were done via web and 24.5% (3291/13,454) of the appointment actions were done via mobile app.

Scheduling Efficiency

The average scheduling actions per finalized visit were similar between self-schedulers and staff schedulers (1.85 average self-scheduled appointment actions per finalized visit vs 2.04 for staff schedulers). There was not a major increase in scheduler work owing to self-scheduling. In fact, staff schedulers averaged only 0.197 appointment actions per finalized visit that had any self-scheduling. Thus, staff rework for patients attempting self-scheduling did not appear to be a major issue. Table 2 also shows that there was a smaller percentage of self-scheduled visits that were cancelled, and many of those were self-cancelled. For the exclusively self-scheduled visit, the appointment process was extremely efficient. Figure 5 shows that 93.5% (7553/8079) of the exclusively self-scheduled patients with a single finalized appointment were able to finalize that appointment in just 1 step (one and done). Thus, only 6.5% (526/8079) of the exclusively self-scheduled patients needed multiple appointment steps for a finalized appointment. However, 25.5% (18,035/70,839) of the staff-scheduled finalized appointments had multiple appointment steps; staff-scheduled

finalized appointments took a single step in 74.5% (52,804/70,839) of the appointment cases. This resulted in an odds ratio of 4.90 (95% CI 4.48-5.37; *P*<.001) for multiple steps in scheduling when comparing staff scheduling to self-scheduling.

Figure 5. Comparison of accumulated percentage of exclusively self-scheduled finalized appointments to that of staff-scheduled finalized appointments by number of appointment steps completed. The graph shows that for each appointment step, the cumulative percentage of self-schedulers successfully completing the appointment process at that step was greater than that of those who used staff schedulers.



Appointment Outcomes: No-shows

For the 12 months studied, there were bilateral screening mammograms scheduled for 93,901unique patients with patient online service access. Of the 131,110 mammograms scheduled, there were 44,146 cancellations, leaving 86,964 scheduled mammograms that were expected to be completed on the scheduled date. Of those appointments expecting to be completed, 95.3% (82,838/86,964) arrived for the visit for an overall no-show rate of 4.7% (4126/86,964) for those with patient online service access. Table 2 shows that the no-show rate for self-scheduled patients was 5.7% (536/9433) compared to 4.6% (3590/77,531) for the staff-scheduled patients. The unadjusted odds ratio of self-scheduled to staff-scheduled no-shows was 1.24 (95% CI 1.13-1.36; *P*<.001). Rosenbaum et al [16] found that patient age was a significant confounder in mammogram no-shows; therefore, we used a multivariable logistic regression model to adjust for age when examining differences in no-shows. In the age-adjusted model, the no-show rates were not significantly different; the age-adjusted odds ratio for self-scheduled to staff-scheduled no-shows was 1.09 (95% CI 0.99-1.20; *P*=.07). However, in a multivariable logistic regression model adjusting for race and ethnicity as well as age, we found a significant no-show odds ratio of self-scheduled to staff-scheduled of 1.12 (95% CI 1.02-1.23; *P*=.02).

Appointment Outcomes: Lead Times

Self-scheduled patients were unable to make their mammogram appointment more than 12 weeks in advance. We found that 4.9% (5778/117,656) of staff-scheduled appointments were scheduled out further than 12 weeks.

Sensitivity Analysis

The percentage of self-scheduled appointment actions is sensitive to the denominator used. Since self-scheduling requires patient online services, we used patients with patient online services as the denominator for our analysis. Portal engagement is not static in many practices. Mayo Clinic patient online services engagement increased from 33% to 62% during 2013 to 2018 [17], and 83.6% (93,901/112,367) of patients scheduling mammograms in our study had patient online services (Figure 3). When including all patients scheduling mammograms (patient online services–enabled or not), there were 151,165 mammograms scheduled over 12 months with a total of 208,521 scheduling and cancelling actions. For the entire cohort of patients with a scheduled mammogram, self-scheduling and

self-cancelling patients performed 9.3% (19,467/208,521) of these actions with staff performing 90.7% (189,054/208,521) of these actions. As shown in Figure 4, the uptake of self-scheduling had increased substantially; therefore, the proportion of self-scheduled actions was also sensitive to the time frame examined. For the last 3 months of the study, the proportion of all self-scheduled mammogram actions (entire cohort of those patient online services–enabled or not) had increased to 12.6% (6109/48,447). The COVID-19 pandemic in spring 2020 resulted in an increase in mammogram cancellations both for self-scheduled and staff-scheduled mammograms. We separately analyzed the 6 pre- and post-COVID months (September 2019 through February 2020 and March 2020 through August 2020, respectively) for the average appointment actions per finalized visit. For self-scheduling, the 6 pre-COVID months had 0.186 staff-scheduling actions per finalized visit compared to 0.209 for the 6 post-COVID months. Thus, even with the COVID-19–associated cancellations, for appointments with any self-scheduling activity, there was only about 1 staff appointment action involved per 5 finalized appointments.

Discussion

Principal Findings

By 11 months, 21.1% (1991/9442) of the patients with self-scheduling access were engaged in self-scheduling their screening mammogram and 24.4% (3285/13,454) of the self-scheduling actions were outside of normal business hours for appointment scheduling. For 93.5% (7553/8079) of those who exclusively self-scheduled their screening mammograms, only 1 appointment step was used—that of a single step of choosing the date and time of the mammogram.

Scheduler Work Implications

Patients performed a large number of scheduling actions, which otherwise would have been done by staff schedulers. There was very little staff-scheduler activity required for each finalized appointment in the self-scheduled group. Thus, there was not an unintended consequence of extra staff-scheduler work required to redo or “clean up” a self-scheduled appointment. We showed that the average self-scheduled finalized appointment involved only 0.197 staff actions compared to 2.04 staff actions on average required for a staff-scheduled finalized appointment. We did not measure the actual staff labor cost for each finalized appointment associated with self-scheduling and staff scheduling. However, with the average self-scheduled finalized visit using only 9.7% (0.197/2.04) of the staff appointment actions compared to a staff-scheduled appointment, there is likely a significant savings. Our findings suggest that the mammogram order generation and self-scheduling features will fit into the cost-effective multicomponent intervention framework for cancer screening identified by Mohan et al [18].

Practice Implications

We did not identify major unintended consequences to the practice. No-shows were significantly greater for those in the self-scheduled group but were reduced to an odds ratio of 1.12 when adjusted for the patient age, race, and ethnicity differences noted in Table 1. Because automated bulk ordering of mammograms was part of the self-scheduling process, providers were freed up to do other activities besides ordering routine mammograms. As preventive services and other chronic care services take up an increasing amount of provider time, decreasing provider time for this activity is very important [8,9].

Patient Implications

Patient self-scheduling is likely a benefit for many patients. We showed that many patients took advantage of the ability to self-schedule 24/7. With 24.5% (3285/13,454) of the self-scheduling occurring after business hours or on weekends and 24.5% (3291/13,454) of the self-scheduling occurring via mobile app, patients were using the anytime and anywhere capability of self-scheduling. Those who self-scheduled also were extremely efficient at doing so, with 93.5% (7553/8079) of their finalized appointments occurring after just 1 scheduling step. Mathioudakis et al [19] noted that women highly value time-efficient screening processes, and our data show the self-scheduling process to be efficient and convenient.

Comparison With Other Studies

There appear to be few comparable studies for self-scheduled imaging. A review of web-based appointment scheduling by Zhao et al [20] focused on medical appointments rather than imaging appointments. Vendors such as Zocdoc or Lybrate offer web-based scheduling of medical appointments but not for imaging [13,21]. Compared to self-scheduled medical appointments, our first year uptake was similar. With a small sample size of 125, Zhang et al [22] found that 11% of patients had used a web-based appointment service for a primary health care center in Australia. We could not find a study like ours comparing no-show mammography appointment outcomes of self-scheduling to staff-scheduling. However, a study by Rosenbaum et al [16] showed a 6.99% no-show for mammography, which is somewhat higher than what we found with either self-scheduled or staff-scheduled mammogram appointments. In Rosenbaum et al’s study, it was noted that younger patients were more likely to no-show their imaging appointments. Given the lower ages in our self-scheduled group, perhaps age was a confounding factor that might explain the higher no-shows in the self-scheduled group. Consistent with Rosenbaum et al’s findings, when we adjusted for age in a multivariable logistic model, there was a nonstatistical difference in no-shows between self-scheduled and staff-scheduled mammogram appointments. However, further adjustment of our no-shows for race and ethnicity as well as age revealed a significant but small association of no-shows with self-scheduling.

The outcomes for self-scheduled mammograms show some interesting contrasts and similarities to outcomes for Mayo Clinic’s self-scheduled well-child visits [15]. Despite differences in patient populations (adult vs pediatric) and appointments scheduled (radiology procedure vs provider visits), there were similar scheduling efficiencies with 93.1% (712/765) of exclusively self-scheduled well-child visits being finalized with 1 appointment step compared to 1 appointment step needed for 93.5% (7553/8079) of the exclusively self-scheduled mammograms. A major difference was in the uptake of

self-scheduling mammograms, which contrasted sharply with that of self-scheduled well-child visits. For the first year of implementation, the percentage of portal-registered unique patients using self-scheduling for mammograms was 15.3% (14,387/93,901) compared to 6.8% (1099/16,161) using well-child visit self-scheduling. An important difference from the well-child appointment process was that self-scheduling a mammogram was paired with a communication process that proactively alerted patients that a mammogram was due. From the user perspective, the electronic mammogram invitation not only notified them that the mammogram was due but also that they could self-schedule their mammogram from their mobile device or online and would not need to phone a scheduler. Although well-child appointments could be self-scheduled, they were not linked to an order that determined eligibility; no proactive well-child appointment due notices (scheduling invitations) were sent out. The pairing of alerting patients to schedule their mammogram appointment and allowing them to self-schedule in a "one stop" process may explain at least some of the two-fold differences in uptake between the mammogram self-scheduling and well-child self-scheduling. Since there were significant differences in population demographics, appointment types, and self-scheduling processes between the mammogram and well-child self-scheduled appointments, more work needs to be done to understand the differences and similarities in the outcomes.

Limitations

Patients self-scheduling mammograms were 93.7% (13,474/14,382) White, and 83.6% (93,901/112,367) of all patients scheduling mammograms were registered with patient online services. Other populations could have different results. Even with our comparison limited to the 83.6% (93,901/112,367) of patients who had patient online service access, there were still significant differences in the ages of patients self-scheduling versus those using staff schedulers. The COVID-19 pandemic occurring in the last 6 months of this study limits some of our findings. However, our subgroup analysis into pre- and post-COVID time frames shows that the extra staff scheduler cancellations due to COVID was associated with only a small increase in average staff-scheduling activity in the self-scheduled group. No-show outcomes in imaging examinations are known to be influenced by a number of factors that we did not take into consideration. For example, in their review of over 3 million outpatient radiology visits, Mieloszyk et al [23] found significant associations of no-shows with patient income, commute distance, and daily snowfall. There is no uniform standard for mammogram screening; there are several somewhat differing recommendations from different specialty organizations and stakeholder groups [2,24,25]. Bitencourt et al [26] discuss some of the differences between breast cancer screening guidelines. Clinics that use different criteria for screening mammography may have different results.

We limited the mammogram self-scheduling feature to a 12-week window as mentioned above. This limits the conclusions about some of the scheduling efficiency. It is possible that the 12-week appointment window resulted in patients having more clarity on their future availability and some reschedules were avoided. Further, the inability to self-schedule more than 12 weeks in the future likely had an impact on the uptake of this feature. Since 4.9% (5778/117,656) of the staff-scheduled appointments were scheduled greater than 12 weeks out, there are likely patients who may have self-scheduled had they had the opportunity to schedule past the 12-week limit. We only examined mammograms that were scheduled. We did not look at the potential issues involved in the identification of individuals who met the criteria for generating a mammogram order. For example, if a mammogram had been recently done elsewhere, the patient might have been misidentified as being due for a mammogram. Further, patients who had changed their email or postal address and had not changed their address in their EHR might not have received their invitation for ordering a mammogram that was due. Our data only reflected those who had acted on screening mammogram orders. In this study, we did not evaluate the accuracy of the mammogram orders or if the patients had received their invitations.

Future Research and Enhancements

Additional research will be needed to evaluate whether web and mobile mammogram self-scheduling will lead to a higher percentage of women receiving timely screening mammograms. A study by Gann et al [27] had an unexpected finding of a greater than 8% increase in mammogram utilization in practices with "active scheduling" compared to "passive scheduling." "Active scheduling" was defined as patients engaged in scheduling their own mammogram, whereas "passive scheduling" was when the clinic actually made the appointment for the patient. Perhaps self-scheduling via web and mobile self-scheduling will be the internet equivalent of "active scheduling" and associated with increased mammogram utilization. Since there are patients who are having mammograms ordered and scheduled greater than 12 weeks in the future, a possible enhancement would be to expand that window of opportunity to self-schedule. A message to the patient noting that a mammogram would be due in 4-6 months and offering a wider window of future times to self-schedule could be an enhancement to evaluate.

Conclusion

A large number of patients successfully self-scheduled their screening mammogram by using the web or mobile without staff-scheduler assistance. Self-scheduling actions were accomplished outside of normal staff-scheduling hours in 24.4% (3285/13,454) of the cases, and 93.5% (7553/8079) of exclusive self-scheduled mammogram appointments were done with just 1 appointment step (one and done). Self-scheduled screening mammograms were associated with more no-shows than staff-scheduled mammograms, with a small but significant odds ratio of 1.12 in a model adjusted for age, race, and ethnicity. There was no unintended consequence of an increase in staff-scheduler work because, on average, each finalized self-scheduled mammogram used less than one-tenth the staff-scheduler appointment actions compared to those completely staff-scheduled.

Authors' Contributions

FN conceived the study, analyzed the data, interpreted it, drafted the manuscript, and performed the statistical analysis. FN, EMN, RJB, RJM, and MCT contributed to the study design. FN, EMN, RJM, RJB, MCT, and BAC contributed to the final manuscript editing, critical revisions, and approval.

Conflicts of Interest

None declared.

References

1. Breast cancer risk in American women. National Cancer Institute. URL: https://www.cancer.gov/types/breast/risk-fact-sheet [accessed 2021-11-25]
2. Recommendations for the early detection of breast cancer. American Cancer Society. URL: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/american-cancer-society-recommendations-for-the-early-detection-of-breast-cancer.html [accessed 2021-11-25]
3. Cancer prevention and early detection facts and figures 2019-2020. American Cancer Society. URL: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/cancer-prevention-and-early-detection-facts-and-figures/cancer-prevention-and-early-detection-facts-and-figures-2019-2020.pdf [accessed 2021-11-25]
4. Duffy SW, Myles JP, Maroni R, Mohammad A. Rapid review of evaluation of interventions to improve participation in cancer screening services. J Med Screen 2017 Sep;24(3):127-145 [FREE Full text] [doi: 10.1177/0969141316664757] [Medline: 27754937]
5. Shires DA, Stange KC, Divine G, Ratliff S, Vashi R, Tai-Seale M, et al. Prioritization of evidence-based preventive health services during periodic health examinations. Am J Prev Med 2012 Feb;42(2):164-173 [FREE Full text] [doi: 10.1016/j.amepre.2011.10.008] [Medline: 22261213]
6. Ridley J, Ischayek A, Dubey V, Iglar K. Adult health checkup: Update on the Preventive Care Checklist Form©. Can Fam Physician 2016 Apr;62(4):307-313 [FREE Full text] [Medline: 27076540]
7. Krogsbøll LT, Jørgensen KJ, Gøtzsche PC. General health checks in adults for reducing morbidity and mortality from disease. Cochrane Database Syst Rev 2019 Jan 31;1:CD009009 [FREE Full text] [doi: 10.1002/14651858.CD009009.pub3] [Medline: 30699470]
8. North F, Tulledge-Scheitel SM, Matulis JC, Pecina JL, Franqueira AM, Johnson SS, et al. Population health challenges in primary care: What are the unfinished tasks and who should do them? SAGE Open Med 2018;6:2050312118800209 [FREE Full text] [doi: 10.1177/2050312118800209] [Medline: 30245819]
9. Yarnall KSH, Pollak KI, Østbye T, Krause KM, Michener JL. Primary care: is there enough time for prevention? Am J Public Health 2003 Apr;93(4):635-641. [doi: 10.2105/ajph.93.4.635] [Medline: 12660210]
10. The power behind 'patient-powered search'. Zocdoc. 2017 Mar 21. URL: https://www.zocdoc.com/about/blog/tech/power-behind-patient-powered-search/ [accessed 2021-11-25]
11. How Zocdoc uses Redis cache to power availability data. Zocdoc. 2019 Nov 11. URL: https://www.zocdoc.com/about/blog/tech/how-zocdoc-uses-redis-cache-to-power-availability-data/ [accessed 2021-11-24]
12. How Zocdoc search works. Zocdoc. URL: https://www.zocdoc.com/about/how-search-works/ [accessed 2021-11-25]
13. Zocdoc.com. URL: https://zocdoc.com [accessed 2021-11-25]
14. Judson TJ, Odisho AY, Neinstein AB, Chao J, Williams A, Miller C, et al. Rapid design and implementation of an integrated patient self-triage and self-scheduling tool for COVID-19. J Am Med Inform Assoc 2020 Jun 01;27(6):860-866 [FREE Full text] [doi: 10.1093/jamia/ocaa051] [Medline: 32267928]
15. North F, Nelson EM, Majerus RJ, Buss RJ, Thompson MC, Crum BA. Impact of Web-Based Self-Scheduling on Finalization of Well-Child Appointments in a Primary Care Setting: Retrospective Comparison Study. JMIR Med Inform 2021 Mar 18;9(3):e23450 [FREE Full text] [doi: 10.2196/23450] [Medline: 33734095]
16. Rosenbaum J, Mieloszyk R, Hall C, Hippe D, Gunn M, Bhargava P. Understanding Why Patients No-Show: Observations of 2.9 Million Outpatient Imaging Visits Over 16 Years. J Am Coll Radiol 2018 Jul;15(7):944-950. [doi: 10.1016/j.jacr.2018.03.053] [Medline: 29755001]
17. North F, Luhman KE, Mallmann EA, Mallmann TJ, Tulledge-Scheitel SM, North EJ, et al. A Retrospective Analysis of Provider-to-Patient Secure Messages: How Much Are They Increasing, Who Is Doing the Work, and Is the Work Happening After Hours? JMIR Med Inform 2020 Jul 08;8(7):e16521 [FREE Full text] [doi: 10.2196/16521] [Medline: 32673238]
18. Mohan G, Chattopadhyay S, Ekwueme D, Sabatino S, Okasako-Schmucker D, Peng Y, Community Preventive Services Task Force. Economics of Multicomponent Interventions to Increase Breast, Cervical, and Colorectal Cancer Screening: A Community Guide Systematic Review. Am J Prev Med 2019 Oct;57(4):557-567 [FREE Full text] [doi: 10.1016/j.amepre.2019.03.006] [Medline: 31477431]
19. Mathioudakis AG, Salakari M, Pylkkanen L, Saz-Parkinson Z, Bramesfeld A, Deandrea S, et al. Systematic review on women's values and preferences concerning breast cancer screening and diagnostic services. Psychooncology 2019 May;28(5):939-947 [FREE Full text] [doi: 10.1002/pon.5041] [Medline: 30812068]

20. Zhao P, Yoo I, Lavoie J, Lavoie BJ, Simoes E. Web-Based Medical Appointment Systems: A Systematic Review. J Med Internet Res 2017 Apr 26;19(4):e134 [FREE Full text] [doi: 10.2196/jmir.6747] [Medline: 28446422]
21. Lybrate.com. URL: https://www.lybrate.com/ [accessed 2021-11-25]
22. Zhang X, Yu P, Yan J. Patients' adoption of the e-appointment scheduling service: A case study in primary healthcare. Stud Health Technol Inform 2014;204:176-181. [Medline: 25087546]
23. Mieloszyk R, Rosenbaum J, Hall C, Hippe D, Gunn M, Bhargava P. Environmental Factors Predictive of No-Show Visits in Radiology: Observations of Three Million Outpatient Imaging Visits Over 16 Years. J Am Coll Radiol 2019 Apr;16(4 Pt B):554-559. [doi: 10.1016/j.jacr.2018.12.046] [Medline: 30947887]
24. Breast cancer: screening. US Preventive Services Task Force. 2016 Jan 11. URL: https://www.uspreventiveservicestaskforce.org/uspstf/recommendation/breast-cancer-screening [accessed 2021-11-26]
25. Qaseem A, Lin JS, Mustafa RA, Horwitch CA, Wilt TJ. Screening for Breast Cancer in Average-Risk Women: A Guidance Statement From the American College of Physicians. Ann Intern Med 2019 Apr 09;170(8):547. [doi: 10.7326/m18-2147]
26. Bitencourt AG, Rossi Saccarelli C, Kuhl C, Morris EA. Breast cancer screening in average-risk women: towards personalized screening. Br J Radiol 2019 Nov;92(1103):20190660 [FREE Full text] [doi: 10.1259/bjr.20190660] [Medline: 31538501]
27. Gann P, Melville SK, Luckmann R. Characteristics of primary care office systems as predictors of mammography utilization. Ann Intern Med 1993 Jun 01;118(11):893-898. [doi: 10.7326/0003-4819-118-11-199306010-00011] [Medline: 8480964]

Abbreviations

EHR: electronic health record

Edited by C Lovis; submitted 08.02.21; peer-reviewed by Y Chu; comments to author 02.03.21; revised version received 03.04.21; accepted 15.11.21; published 07.12.21.

Please cite as:
North F, Nelson EM, Buss RJ, Majerus RJ, Thompson MC, Crum BA
The Effect of Automated Mammogram Orders Paired With Electronic Invitations to Self-schedule on Mammogram Scheduling Outcomes: Observational Cohort Comparison
JMIR Med Inform 2021;9(12):e27072
URL: https://medinform.jmir.org/2021/12/e27072
doi:10.2196/27072
PMID:34878997

Original Paper

Leveraging National Claims and Hospital Big Data: Cohort Study on a Statin-Drug Interaction Use Case

Aurélie Bannay[1,2], MD, PhD; Mathilde Bories[2,3,4], MSc; Pascal Le Corre[3,4,5], PharmD, PhD; Christine Riou[2], MD; Pierre Lemordant[2], ME; Pascal Van Hille[2], ME; Emmanuel Chazard[6], MD, PhD; Xavier Dode[7,8], PharmD; Marc Cuggia[2], MD, PhD; Guillaume Bouzillé[2], MD, PhD

[1]Université de Lorraine, Centre Hospitalier Régional Universitaire de Nancy, Centre national de la recherche scientifique, Inria, Laboratoire lorrain de recherche en informatique et ses applications, Nancy, France

[2]Inserm, Laboratoire Traitement du Signal et de l'Image - UMR 1099, Centre Hospitalier Universitaire de Rennes, Université de Rennes 1, Rennes, France

[3]Pôle Pharmacie, Service Hospitalo-Universitaire de Pharmacie, Centre Hospitalier Universitaire de Rennes, Rennes, France

[4]Laboratoire de Biopharmacie et Pharmacie Clinique, Faculté de Pharmacie, Université de Rennes 1, Rennes, France

[5]Centre Hospitalier Universitaire de Rennes, Inserm, Ecole des hautes études en santé publique, Institut de recherche en santé, environnement et travail, UMR_S 1085, Université de Rennes 1, Rennes, France

[6]Centre d'Etudes et de Recherche en Informatique Médicale EA2694, Centre Hospitalier Universitaire de Lille, Université de Lille, Lille, France

[7]Centre National Hospitalier d'Information sur le Médicament, Paris, France

[8]Department of Pharmacy, Hospices Civils de Lyon, University Hospital, Lyon, France

Corresponding Author:
Guillaume Bouzillé, MD, PhD
Inserm, Laboratoire Traitement du Signal et de l'Image - UMR 1099
Centre Hospitalier Universitaire de Rennes
Université de Rennes 1
UFR Santé, laboratoire d'informatique médicale
2 avenue du Professeur Léon Bernard
Rennes, 35000
France
Phone: 33 615711230
Email: guillaume.bouzille@gmail.com

Abstract

Background: Linking different sources of medical data is a promising approach to analyze care trajectories. The aim of the INSHARE (Integrating and Sharing Health Big Data for Research) project was to provide the blueprint for a technological platform that facilitates integration, sharing, and reuse of data from 2 sources: the clinical data warehouse (CDW) of the Rennes academic hospital, called eHOP (entrepôt Hôpital), and a data set extracted from the French national claim data warehouse (Système National des Données de Santé [SNDS]).

Objective: This study aims to demonstrate how the INSHARE platform can support big data analytic tasks in the health field using a pharmacovigilance use case based on statin consumption and statin-drug interactions.

Methods: A Spark distributed cluster-computing framework was used for the record linkage procedure and all analyses. A semideterministic record linkage method based on the common variables between the chosen data sources was developed to identify all patients discharged after at least one hospital stay at the Rennes academic hospital between 2015 and 2017. The use-case study focused on a cohort of patients treated with statins prescribed by their general practitioner or during their hospital stay.

Results: The whole process (record linkage procedure and use-case analyses) required 88 minutes. Of the 161,532 and 164,316 patients from the SNDS and eHOP CDW data sets, respectively, 159,495 patients were successfully linked (98.74% and 97.07% of patients from SNDS and eHOP CDW, respectively). Of the 16,806 patients with at least one statin delivery, 8293 patients started the consumption before and continued during the hospital stay, 6382 patients stopped statin consumption at hospital admission, and 2131 patients initiated statins in hospital. Statin-drug interactions occurred more frequently during hospitalization than in the community (3800/10,424, 36.45% and 3253/14,675, 22.17%, respectively; $P<.001$). Only 121 patients had the most

severe level of statin-drug interaction. Hospital stay burden (length of stay and in-hospital mortality) was more severe in patients with statin-drug interactions during hospitalization.

Conclusions: This study demonstrates the added value of combining and reusing clinical and claim data to provide large-scale measures of drug-drug interaction prevalence and care pathways outside hospitals. It builds a path to move the current health care system toward a Learning Health System using knowledge generated from research on real-world health data.

(JMIR Med Inform 2021;9(12):e29286) doi:10.2196/29286

KEYWORDS

drug interactions; statins; administrative claims; health care; big data; data linking; data warehousing

Introduction

The secondary use of health care data offers the opportunity to conduct observational studies in real life [1-3]. Indeed, hospital clinical data warehouses (CDWs) supply fine-grained information from electronic health records (EHRs), such as laboratory test results and drug administration, but are restricted to hospitalized patients. Conversely, National claim databases offer limited information (eg, drug reimbursement and health care consumption data), but on a large part of the population. Therefore, matching the data from these 2 different databases could be informative, but it is also challenging. Patients existing in the 2 databases should be correctly identified using appropriate record linkage methods. The first option is deterministic record linkage that relies on the presence of a unique common identifier or a combination of different variables used as a key to join tables [4]. More complex rules to link records can also be added, such as an acceptable distance between string variables or between dates. The second option is probabilistic record linkage that is based on a model to assess the discriminative power of each variable used in the record linkage strategy. The result is the probability that an entity in the first database is the same entity in the second database [5,6]. Several studies have demonstrated that in most cases, probabilistic approaches give better results than deterministic methods [7-10]. However, the choice of record linkage also heavily depends on the characteristics of the 2 databases to be linked. The quality of the data used in the record linkage is an especially important factor. Indeed, if high quality data (eg, few missing values) are available, deterministic methods can achieve good results and are easier to develop [11].

In France, the national health database, Système National des Données de Santé (SNDS), [12] links the nationwide outpatient claim database (Système national d'information inter-régimes de l'Assurance maladie), the national discharge database (Programme de Médicalisation des Systèmes d'Information [PMSI]), and the Epidemiology Centre of Medical Causes of Death (CepiDC; vital status data) database. Rennes academic hospital (Centre Hospitalier Universitaire de Rennes) uses eHOP (entrepôt Hôpital) [13], a CDW that includes EHR and discharge data on all stays in this hospital. Linking SNDS and eHOP is a promising strategy to analyze patient care trajectories. However, legal, methodological, and technical barriers still remain. Health data are sensitive, and in France, their use is regulated by the European General Data Protection Regulation [14]. Therefore, studies based on the use of health data entail various regulatory steps, such as the scientific evaluation of the project and the patient information material and the assessment of the impact on data protection. In France, the use of SNDS data for external research requires the development of a data repository that complies with the strict security specifications to host the SNDS sample for the study.

In this context, the aim of the INSHARE (Integrating and Sharing Health Big Data for Research) project was to provide the blueprint for a technological platform (INSHARE platform) that facilitates data integration, sharing and reuse by following the FAIR (findability, accessibility, interoperability, and reusability) Guiding Principles [15]. This work demonstrates through a use case in pharmacovigilance how the INSHARE platform can support health big data analysis.

Our use case focused on statin consumption and statin-related drug–drug interactions (DDIs). Indeed, 36.9% [16] of French people aged 34 to 65 years have hypercholesterolemia, and statins are the most prescribed lipid-lowering treatment drugs in France [17]. The current European treatment guidelines [18] recommend statins as the first-choice drug for hypercholesterolemia management. However, 10% to 25% of patients treated with statins experience muscle side effects [19], including rhabdomyolysis (incidence: 1-3 in 100,000 persons per year) [20]. Statin-induced rhabdomyolysis is related to DDIs in 60% of cases [20], which suggests that avoiding DDIs has an important role in reducing statin adverse events. Because of their wide use and DDI potential, statins are an interesting study topic to assess the value of our technological platform for clinical data reuse. Moreover, literature data indicate that DDIs are preventable, but this is hindered by the clinicians' lack of easy access to comprehensive information. Indeed, health care delivery is fragmented across the system and this creates an environment susceptible to medication-related issues [21]. Polypharmacy has been associated with higher risk of DDIs and adverse drug events [22], and subsequently, with drug-related deaths in hospitals [23]. Therefore, it is important to precisely characterize the individual care pathways within the health care system using aggregated medical data.

Here, we present the technical aspects of the INSHARE platform and the methods and results of the care pathway analysis in patients with statin-drug interactions.

Methods

Data Sources

Drug Database: Thériaque

Thériaque is a comprehensive dynamic knowledge database that provides exhaustive information on approved and marketed drugs [24]. It contains highly structured information on each drug, such as indications, contraindications, and DDIs and their severity level. Each drug is referenced according to 3 mapped classifications: Unité Commune de Dispensation, the medication-dispensing unit used by the French hospital information system; Code Identifiant de la Présentation, the drug package identifier used by French community pharmacies; and Anatomical Therapeutic Codification, which is based on the active component or components of each drug.

French Claim Database: SNDS

In France, the SNDS is a national claim data warehouse that covers 98.8% of the entire French population [25]. It contains data from outpatient care, such as medical consultations and drug deliveries by community pharmacies, and data from inpatient care, such as diagnosis and procedures performed during a stay in a private- or public-sector hospital. Each reimbursement of outpatient care is recorded at the individual level in a specific data mart called Datamart de Consommation InterRégime [12]. Data on inpatient care also are recorded at the individual level in an annual national discharge database called PMSI that is similar to the diagnosis-related groups. Individual data are deidentified and pseudonymized allowing the linkage, thanks to a unique identifier, between inpatient data (PMSI database) and outpatient data (Datamart de Consommation Inter Régime). This claim data warehouse has been previously described [12].

We used a data set extracted from the SNDS database that included all patients discharged after at least one hospital stay at Rennes academic hospital between 2015 and 2017. Owing to the redundancy of information contained in the PMSI database, hospital stays following the primary diagnosis were excluded (eg, stays for chemotherapy, radiotherapy, dialysis, apheresis, blood transfusion and hyperbaric oxygen therapy). All inpatient and outpatient data in the 12 months before each hospital stay were extracted.

Data were extracted from the national SNDS database by a French national health insurance manager outside of this study workflow.

CDW: eHOP

eHOP is the CDW developed and deployed at Rennes academic hospital [13]. It collects administrative and clinical data from EHRs, both unstructured (eg, clinical notes) and structured (eg, drugs, laboratory results). Data are deidentified and a unique anonymous identifier allows the linkage among hospital stays of a given patient. The eHOP CDW currently allows for searching from 80 million unstructured data and 430 million structured elements. All these data are collected from EHRs and cover more than 1.4 million patients.

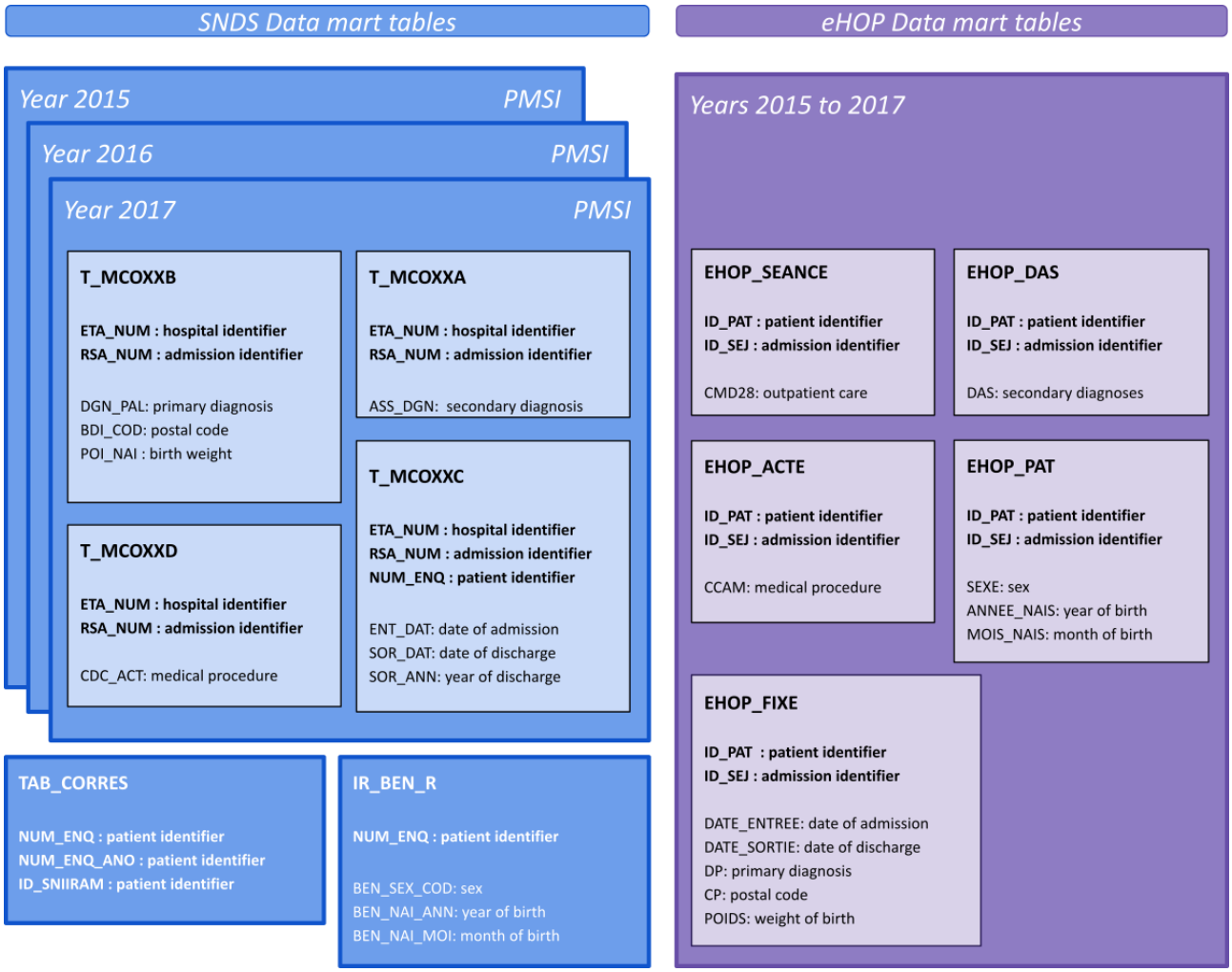
The data set from the eHOP database included patients according to the same criteria used for the SNDS data: all data on hospital stays at Rennes academic hospital between 2015 and 2017. For this study, we used the following structured data:

1. Demographic data
2. Drug administered (Common Unit of Dispensation, UCD and date of administration)
3. PMSI data: *International Classification of Diseases, Tenth Revision (ICD-10)* codes, procedure codes, mortality, length of stay, etc.
4. Laboratory results described with a local terminology.

Record Linkage Procedure

As no unique patient identifier is available to link SNDS and eHOP data because of regulatory issues, we developed a semideterministic record linkage method based on PMSI variables that are common between the SNDS data source and the eHOP CDW data source (Figure 1). PMSI data are available from all French hospitals and are produced in a standardized way by each hospital. Once deidentified, PMSI data feed the nationwide PMSI database. This database is then integrated in the SNDS database to link PMSI data with claim data. In theory, PMSI data from the SNDS and hospitals should be exactly the same. However, during the preliminary work, we identified some discrepancies concerning *ICD-10* and procedure codes between these data sources. Therefore, we incorporated some fuzzy logic in the record linkage algorithm to solve inconsistencies. The algorithm is illustrated in Figure 2. Specifically, *ICD-10* codes comprise between 3 and 6 characters, but we kept only the first 4 characters. Procedure codes comprise 7 characters, and we kept all 7. We merged *ICD-10* and procedure codes in alphabetical order in a unique string for each stay. We then tested different Levenshtein distance thresholds to consider a match between sets of codes (the distribution of the Levenshtein distances for the *ICD-10* codes and procedure codes is provided in Multimedia Appendix 1, Table S1). We identified a threshold of 5 as the best choice for both *ICD-10* and procedure codes. For the final matching, first we assessed whether a patient had at least one exact match. This was considered as the exact match if the other patients were fuzzy matches. If we did not find any exact match, we kept the fuzzy match first looking at procedure codes. If a patient had several exact matches or several fuzzy matches, we kept the one with the most fuzzy matches on *ICD-10* codes. The remaining patients with several matches were considered as duplicates and were excluded from the linkage results.

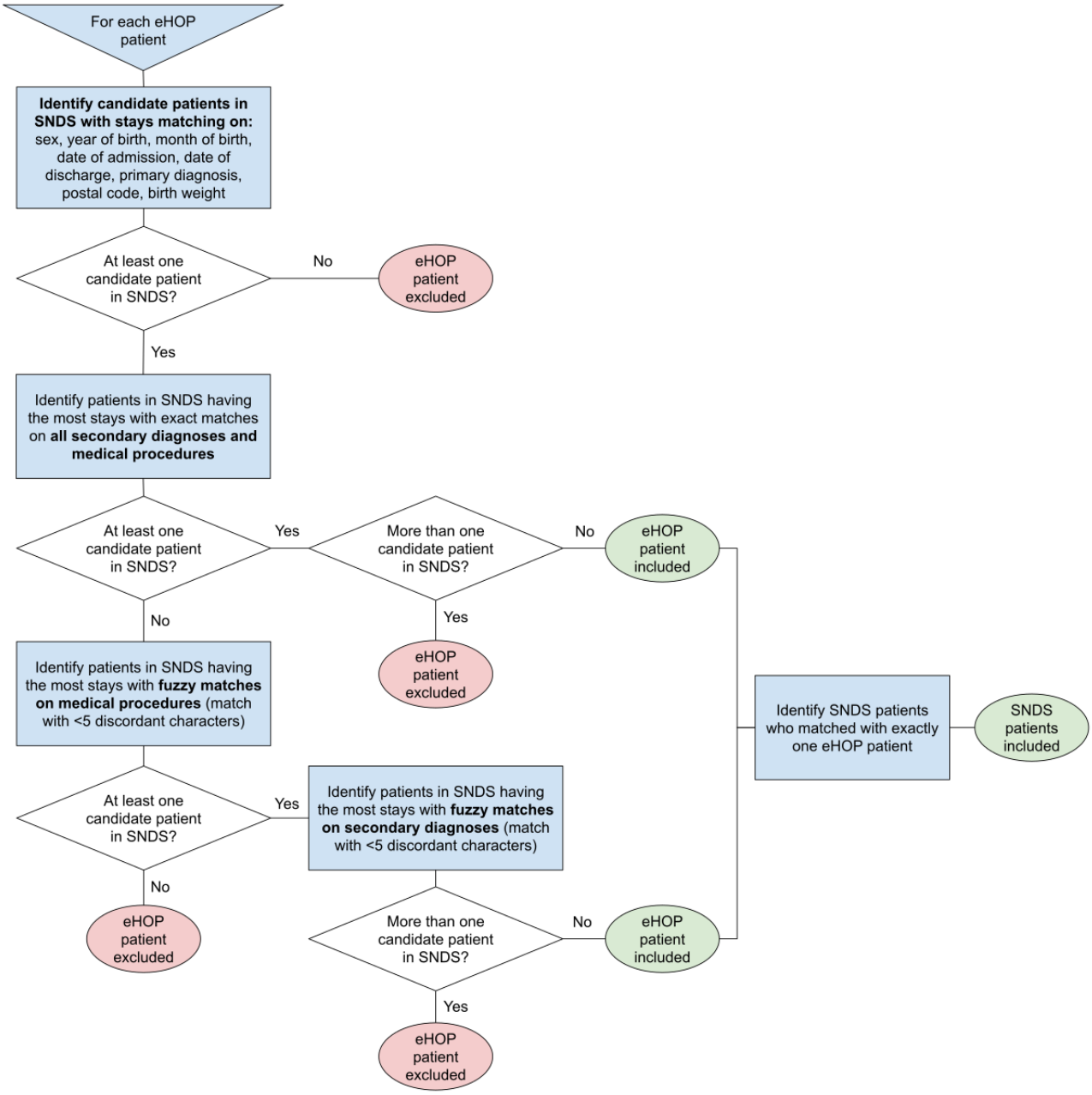
Figure 1. SNDS data mart tables (in blue), including PMSI tables, and eHOP data mart tables (in purple) with the different variables from the 2 data sources used for the linkage procedure. eHOP: entrepôt Hôpital; PMSI: Programme de Médicalisation des Systèmes d'Information; SNDS: Système National des Données de Santé.



We also had to solve specific cases concerning twins who do not have an individual identifier (NUM_ENQ) in the PMSI. Indeed, the same identifier (NUM_ENQ) is shared by twins of the same sex [12]. Thus, it was impossible to link their SNDS records with their records in eHOP. We chose to exclude twin patients from the record linkage results. The complete algorithm is available in Multimedia Appendix 1, Figure S1.

We assessed the linkage effectiveness by calculating the rate of SNDS and eHOP patients who could be matched in the other data set. We also describe some characteristics of the following groups: patients who were matched between data sources, and patients from the SNDS and eHOP data sets who could not be matched.

Figure 2. Decision tree for the record linkage procedure. eHOP: entrepôt Hôpital; SNDS: Système National des Données de Santé.



INSHARE Platform

The INSHARE platform comprises 2 parts: a data repository to gather all kinds of data sources, and a computing infrastructure to perform data preparation, record linkage and analyses. The platform is available through Apache Mesos, a resource manager, to allow concurrent access to the computing server.

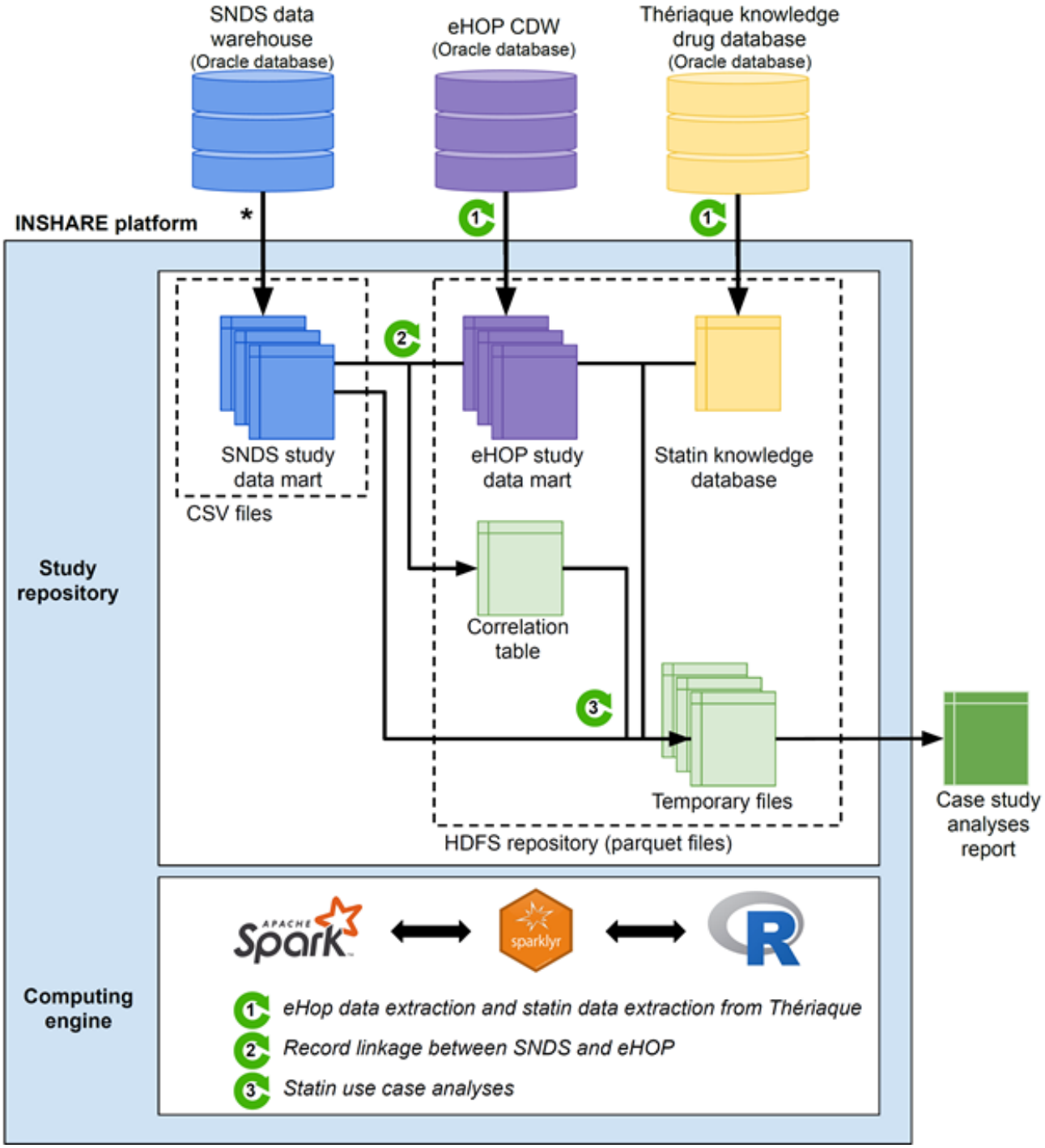
The data repository was the Apache Hadoop Distributed File System (HDFS) repository, and data were stored in parquet format files, with an appropriate stratification key. SNDS data sets were made available to us in CSV files that were stored in a specific folder in the server. We extracted the data needed from the eHOP CDW and the Thériaque databases with Spark SQL. This extraction step avoided repeating long queries in the CDW and overloading the production CDW used for other purposes.

We used the Spark distributed computing framework, version 2.3.4, for the data preparation, the record linkage procedure, and all use-case analyses.

We then accessed these data with Spark SQL that allowed us to merge data from the different sources in an efficient way and to perform all analyses. We used the R language as the script language, particularly the sparklyr package. The overall data processing is depicted in Figure 3.

We used a single node cluster: a CentOS 7 Unix server with 2 Intel Xeon 5122@3.6 GHz and 192 GB of RAM. Thus, we did not replicate the HDFS repository, and we executed the Spark master and slave nodes on the same machine.

Figure 3. INSHARE platform and data processing workflow. CDW: clinical data warehouse; eHOP: entrepôt Hôpital; HDFS: Hadoop Distributed File System; INSHARE: Integrating and Sharing Health Big Data for Research; SNDS: Système National des Données de Santé.



Use Case Study Design

We performed a cohort study on patients treated with statins prescribed by their general practitioners or during the hospital stay. We collected information on statins (Anatomical Therapeutic Codification classes C10AA, C10BA, and C10BX) and the statin-drug interactions from the Thériaque database. We classified statin intake as (1) community consumption if we found at least one statin delivery by a community pharmacy less than 1 month before hospitalization, and (2) hospital consumption if we found at least one statine administered during the hospital stay. Only the first hospital stay for each patient was retained for the use-case.

For each patient, we extracted the following features: sex, age at admission, the international nonproprietary name of the used statin, consumption of drugs potentially interacting with the used statin, DDI severity, admission via the emergency department, length of hospital stay, in-hospital death, laboratory results: creatine phosphokinase (CPK), creatinineaemia, glycemia, hemoglobin, kalemia, natremia, aspartate aminotransferase, alanine aminotransferase, hospital care burden (ie, diagnosis-related group severity).

We classified patients into 3 subgroups according to their statin consumption status: (1) patients treated with statins before and during their hospital stay, (2) patients treated with statins before admission, but not during the hospital stay, (3) patient who started taking statins in hospital without any statin treatment in the previous 12 months. We defined a statin-related DDI on the basis of the intake of a drug that reacts with the statin taken by that patient. All hospital drug administrations were considered during the index hospital stay, and all community deliveries were considered within 8 days before or after the statin delivery. According to the Thériaque database, we classified all statin-drug interactions into 3 levels of severity (level 1: contraindication, level 2: relative contraindication, level 3: precaution of use).

Statistical Analyses

We described categorical variables as numbers and percentages, and quantitative variables as mean and SD for symmetrical distribution, and median with first and third quartiles (Q1–Q3), otherwise. We explored the association between patient characteristics or hospital stays and the occurrence of a statin-drug interaction with the Chi-square test (categorical variables) and one-way analysis of variance (quantitative variables). We built a logistic regression model to identify factors independently related to the occurrence of a statin interaction.

Ethical Consideration

The record linkage and the use-case study were approved by the Commission nationale de l'informatique et des libertés (French Data Protection Agency or CNIL; N 2,206,739). According to French regulations, patients were informed about the use of their data, and no signed consent was required.

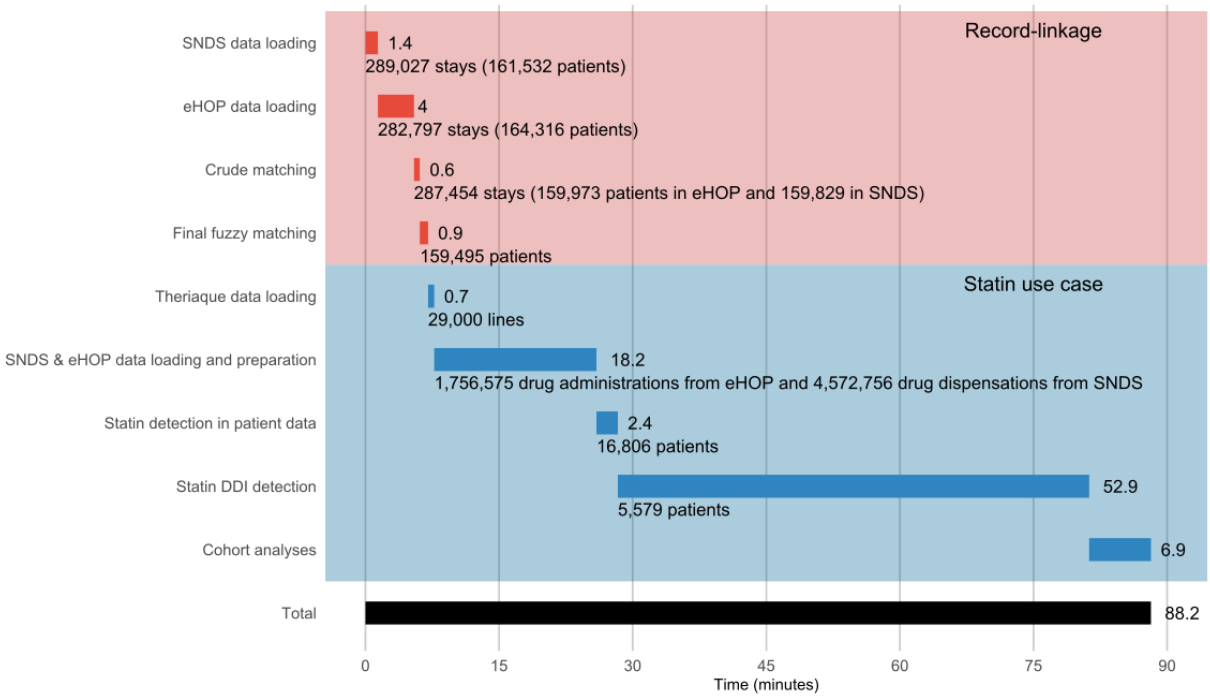
Results

Technological Results

INSHARE Overall Computing Performance

The time needed for the record linkage procedure and statin use-case analysis was 88 minutes. The most time-consuming step was the detection of DDIs in the data of patients taking a statin. The time needed for each step is indicated in Figure 4.

Figure 4. Time duration from data loading to the end of the use case-study analysis. DDI: drug–drug interaction; eHOP: entrepôt Hôpital; SNDS: Système National des Données de Santé.



Assessment of the Record Linkage Procedure

The SNDS and eHOP data sets included 161,532 subjects (278,341 stays) and 164,316 subjects (265,089 stays), respectively, who had at least one hospital stay at Rennes academic hospital between 2015 and 2017.

We successfully linked 159,495 patients (159,495/161,532, 98.74% and 159,495/164,316, 97.07% patients from the SNDS and eHOP data sets, respectively). We excluded from the linkage results 199 patients from the SNDS data set and 162 patients from the eHOP data set because their records were linked with more than one patient in the other data set. Patients who could not be linked were younger (median age of the unmatched patients from the eHOP and SNDS data sets: 22.3 and 27.6 years, respectively, compared with 48.4 years for matched patients). Moreover, women represented 51.35% (81,900/159,495) of all matched patients and 57.20% (2758/4821) and 18.52% (377/2037) of unmatched patients in the eHOP and SNDS data sets, respectively.

Use Case Results

Statin-Taking Population

Of the 159,495 matched patients, we retained 16,806 patients with at least one statine delivery. Specifically, 8293 patients

started statin treatment before admission and continued it during the hospital stay (community and hospital consumption), 6382 patients started statin treatment before admission but stopped at hospital admission (only community consumption), and 2131 patients started statins in the hospital (hospital initiation). The characteristics of the 3 subgroups are summarized in Table 1. Age (4651/6382, 72.88% and 6255/8293, 75.43% of patients aged ≥65 years) and unplanned hospitalization rate (2416/6382, 37.86% and 2434/8293, 29.35%) were similar in patients with only community consumption and patients with community and hospital consumption, respectively. Type of hospital care was similar in patients with community and hospital consumption and in patients with hospital initiation (4729/8293, 57.02% and 1072/2131, 50.31% of surgery, respectively). The percentage of patients aged ≥65 years and the rate of planned hospitalizations were lower in patients with hospital initiation than in the other 2 subgroups.

The most dispensed statin in all 3 subgroups was atorvastatin. Simvastatin, rosuvastatin and pravastatin each represented approximately 1 out of 5 prescriptions in patients with only community consumption. In the hospital, only 2 statins were available (atorvastatin and pravastatin).

Table 1. Patients' characteristics according to their statin consumption.

	Only community consumption (n=6382), n (%)	Community and hospital consumption (n=8293), n (%)	Hospital initiation (n=2131), n (%)	*P* value
Sex (male)	3790 (59.39)	5431 (65.49)	1437 (67.43)	<.001
Age (≥65 years)	4651 (72.88)	6255 (75.43)	1192 (55.94)	<.001
Unscheduled admission	2416 (37.86)	2434 (29.35)	1155 (54.2)	<.001
Type of care				<.001
Medical care	4576 (71.7)	3564 (42.98)	1059 (49.69)	
Surgery	1806 (28.29)	4729 (57.02)	1072 (50.31)	
Chronic statin consumption (>3 months)	6075 (95.19)	7660 (92.37)	—[a]	<.001
Statin type				<.001
Atorvastatin	2380 (37.29)	4632 (55.85)	1909 (89.58)	
Fluvastatin	194 (3.04)	190 (2.29)	3 (0.14)	
Pravastatin	1374 (21.53)	2004 (24.16)	183 (8.58)	
Rosuvastatin	1145 (17.94)	1473 (17.76)	24 (1.13)	
Simvastatin	1301 (20.39)	1540 (18.57)	24 (1.13)	
Patients with statin-drug interactions				
During community consumption	1438 (22.53)	1815 (21.89)	—	<.001
DDI[b] severity				.07
1	30 (2.09)	20 (1.10)	—	
2	20 (1.39)	29 (1.60)	—	
3	1404 (97.64)	1784 (98.29)	—	
During hospital consumption	—	3215 (38.77)	585 (27.45)	<.001
DDI severity				<.001
1	—	72 (2.24)	10 (1.71)	
2	—	143 (4.45)	58 (9.91)	
3	—	3154 (98.10)	552 (94.36)	

[a]Not available.

[b]DDI: drug–drug interaction.

Statin-Drug Interaction Detection

We identified 5579 patients with potential statin-related DDIs. Overall, statin-drug interactions occurred more frequently during hospitalization than in the community (3800/10,424, 36.45% and 3253/14,675, 22.17%, respectively). The most severe DDIs (level 1) concerned 0.78% (82/10,424) of hospitalized patients.

Table 2 presents the hospital outcomes in patients with and without statin-drug interactions. Patients with statin-drug interactions were divided into 3 subgroups according to the place of DDI occurrence: (1) during community consumption (regardless of their hospital consumption), (2) during hospital consumption (regardless of their community consumption), or (3) during both community and hospital consumption. Statin-drug interactions occurring in hospital were associated

with longer hospital stay, more severe pathology, and higher in-hospital mortality. The logistic regression model identified characteristics that were significantly related to the occurrence of statin-drug interactions: men older than 64 years of age, admitted for medical care for severe pathology, and longer length of hospital stay (Table 3).

Tables 4 and 5 present the frequency of patients according to their DDI severity and to the place of DDI occurrence and the details of the 5 most frequent drugs that interacted with statins according to the place of DDI occurrence.

Table 2. Characteristics of patients and hospital stays according to the place of the statin-drug interaction occurrence.

	Interaction only during community consumption (n=1779)	Interaction only during hospital consumption (n=2326)	Interaction during community and also hospital consumption (n=1474)	No interaction (n=11,227)	*P* value
Sex (men), n (%)	1132 (63.63)	1521 (65.39)	1008 (68.39)	6997 (62.32)	<.001
Age (≥65 years), n (%)	1394 (78.36)	1750 (75.24)	1215 (82.43)	7739 (68.93)	<.001
Unscheduled admission, n (%)	698 (39.24)	926 (39.81)	544 (36.91)	3837 (34.18)	<.001
Type of care, n (%)					<.001
Medical care	1233 (69.31)	1149 (49.39)	787 (53.39)	6030 (53.71)	
Surgery	546 (30.69)	1177 (50.6)	687 (46.61)	5197 (46.29)	
Length of stay (days), mean (SD)	8.3 (10.4)	11.9 (15.8)	8.4 (10.1)	7.6 (8.2)	<.001
Intensive care unit admission, n (%)	94 (5.28)	742 (31.9)	260 (17.64)	2156 (19.2)	<.001
Diagnosis-related group severity, n (%)					<.001
1 (least severe)	1314 (73.86)	692 (29.75)	565 (38.33)	7070 (62.97)	
2	177 (9.95)	673 (28.93)	388 (26.32)	2015 (17.95)	
3	197 (11.07)	729 (31.34)	396 (26.87)	1692 (15.07)	
4 (most severe)	91 (5.12)	232 (9.97)	125 (8.48)	450 (4)	
In-hospital mortality, n (%)	12 (0.67)	24 (1.03)	31 (2.1)	87 (0.77)	<.001

Table 3. Factors related to the occurrence of a statin interaction.

	Odds ratio (95% CI)
Sex (male)	1.14 (1.04-1.25)
Age (≥65 years)	1.48 (1.34-1.62)
Unscheduled admission	1.08 (0.97-1.19)
Medical care	1.56 (1.43-1.71)
Length of stay (days)	1.03 (1.03-1.04)
Diagnosis-related group severity	
1 (least severe)	1
2	1.18 (1.06-1.31)
3	1.27 (1.13-1.43)
4 (most severe)	1.51 (1.22-1.86)

Table 4. Top 5 drugs interacting with statins during community consumption, along with the overall total for each security level.

Drug or statin	Rosuvastatin, n (%)	Simvastatin, n (%)	Atorvastatin, n (%)	Pravastatin, n (%)	Fluvastatin, n (%)
Severity level: 1 (most severe)					
Cyclosporin (n=29)	17 (68)	12 (63.2)	0 (0)	0 (0)	0 (0)
Sodium fusidate (n=8)	1 (4)	3 (15.8)	2 (50)	2 (100)	0 (0)
Fenofibrate (n=6)	6 (24)	0 (0)	0 (0)	0 (0)	0 (0)
Telithromycin (n=3)	0 (0)	1 (5.3)	2 (50)	0 (0)	0 (0)
Clarithromycin (n=3)	0 (0)	3 (15.8)	0 (0)	0 (0)	0 (0)
Total	25 (100)	19 (100)	4 (100)	2 (100)	0 (0)
Severity level: 2					
Fenofibrate (n=23)	6 (85.7)	3 (13)	6 (54.5)	3 (100)	5 (83.3)
Carbamazepine (n=15)	0 (0)	15 (65.2)	0 (0)	0 (0)	0 (0)
Cyclosporin (n=6)	0 (0)	4 (17.4)	2 (18.2)	0 (0)	0 (0)
Rifampicin (n=2)	0 (0)	1 (4.3)	1 (9.1)	0 (0)	0 (0)
Bezafibrate (n=2)	1 (14.3)	0 (0)	0 (0)	0 (0)	1 (16.7)
Total	7 (100)	23 (100)	11 (100)	3 (100)	6 (100)
Severity level: 3 (least severe)					
Fluindione (n=898)	156 (26.5)	164 (13.1)	365 (22.2)	194 (25.7)	19 (22.9)
Warfarin sodium (n=674)	114 (19.4)	101 (8)	287 (17.5)	152 (20.1)	20 (24.1)
Amlodipine besylate (n=322)	0 (0)	322 (25.6)	0 (0)	0 (0)	0 (0)
Sodium bicarbonate or sodium alginate[a] (n=285)	51 (8.7)	58 (4.6)	110 (6.7)	57 (7.5)	9 (10.8)
Sodium polystyrene sulfonate (n=225)	38 (6.5)	0 (0)	102 (6.2)	52 (6.9)	0 (0)
Total	588 (100)	1256 (100)	1643 (100)	755 (100)	83 (100)

[a]Sodium bicarbonate-containing antacid.

Table 5. Top 5 drugs interacting with statins during hospital consumption, along with the overall total for each security level.

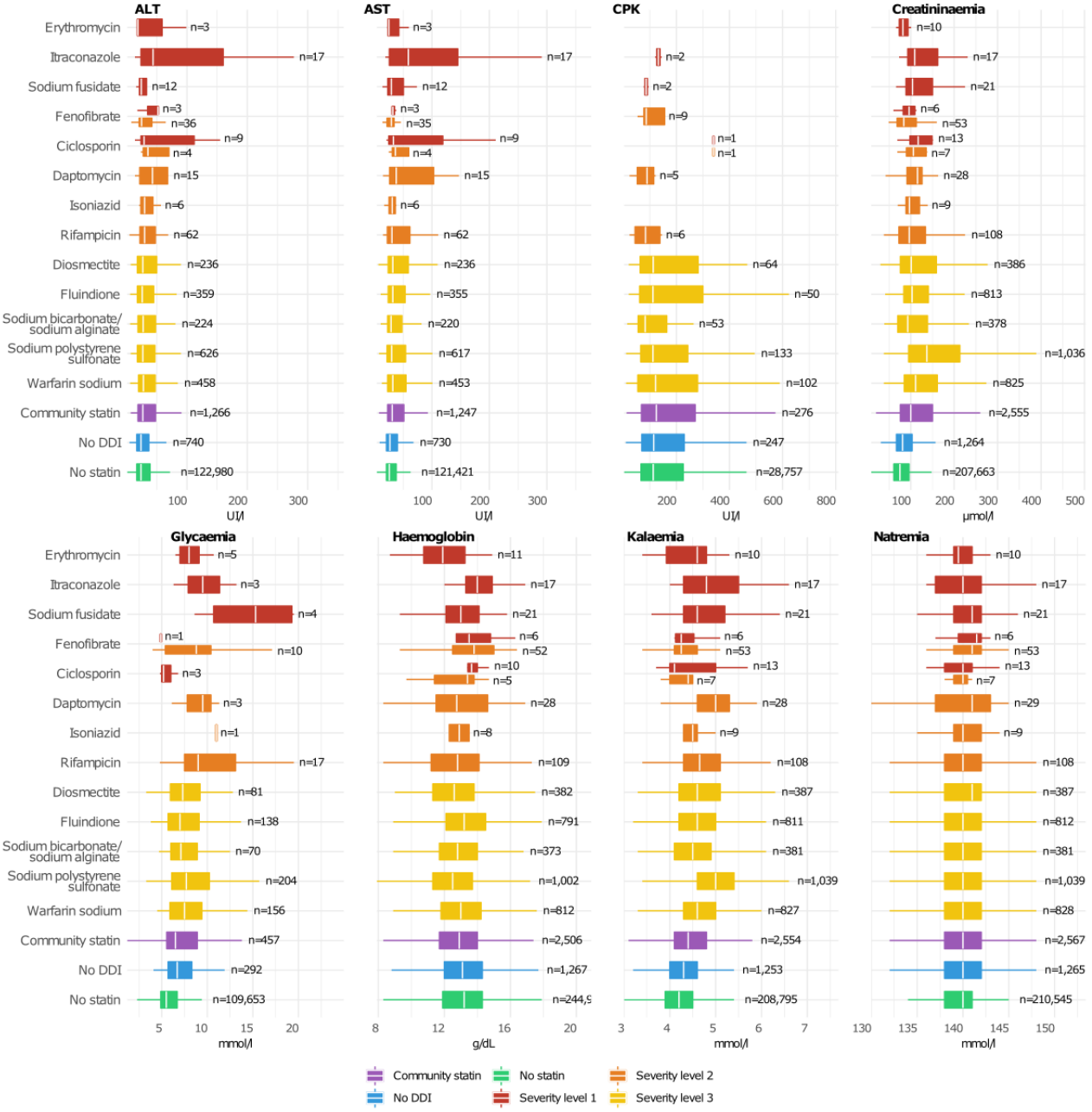
Drug or statin	Rosuvastatin, n (%)	Simvastatin, n (%)	Atorvastatin, n (%)	Pravastatin, n (%)	Fluvastatin, n (%)
Severity level: 1 (most severe)					
Sodium fusidate (n=21)	3 (18.7)	0 (0)	14 (41.2)	4 (80)	0 (0)
Itraconazole (n=19)	0 (0)	2 (6.9)	17 (50)	0 (0)	0 (0)
Cyclosporin (n=15)	6 (37.5)	9 (31)	0 (0)	0 (0)	0 (0)
Erythromycin (n=12)	0 (0)	12 (41.4)	0 (0)	0 (0)	0 (0)
Fenofibrate (n=7)	7 (43.8)	0 (0)	0 (0)	0 (0)	0 (0)
Total	16 (100)	29 (100)	34 (100)	5 (100)	0 (0)
Severity level: 2					
Rifampicin (n=118)	0 (0)	11 (40.7)	107 (56.9)	0 (0)	0 (0)
Fenofibrate (n=56)	7 (58.3)	3 (11.1)	36 (19.1)	8 (61.5)	2 (100)
Daptomycin (n=30)	5 (41.7)	3 (11.1)	19 (10.1)	3 (23.1)	0 (0)
Isoniazid (n=9)	0 (0)	0 (0)	9 (4.8)	0 (0)	0 (0)
Cyclosporin (n=8)	0 (0)	5 (18.5)	3 (1.6)	0 (0)	0 (0)
Total	12 (100)	27 (100)	188 (100)	13 (100)	2 (100)
Severity level: 3 (least severe)					
Sodium polystyrene sulfonate (n=1142)	100 (19.7)	98 (9.4)	660 (18.7)	243 (20.2)	13 (16.5)
Warfarin sodium (n=894)	83 (16.3)	65 (6.2)	529 (14.9)	206 (17.1)	11 (13.9)
Fluindione (n=894)	92 (18.1)	88 (8.5)	504 (14.3)	196 (16.3)	14 (17.7)
Diosmectite (n=431)	42 (8.3)	31 (2.9)	259 (7.3)	93 (7.7)	6 (7.6)
Sodium bicarbonate or sodium alginate[a] (n=421)	39 (7.7)	39 (3.7)	242 (6.8)	93 (7.7)	8 (10.1)
Total	508 (100)	1040 (100)	3531 (100)	1204 (100)	79 (100)

[a]Sodium bicarbonate-containing antacid.

Link Between Statin-Drug Interaction and Laboratory Results

Figure 5 illustrates the link between the 5 most frequent drug interactions of each statin and the laboratory results. Overall, we observed little variations in laboratory values between patients with level 3 statin-drug interactions and patients without statins or taking statins but without DDI. However, glycemia was higher in patients in whom a potential statin interaction (level 1) with sodium fusidate, itraconazole, or erythromycin was detected. Similarly, kalemia and liver enzymes (alanine aminotransferase, aspartate aminotransferase) were altered in patients with a potential statin interaction (level 1) with itraconazole, or sodium fusidate, and with itraconazole, respectively. However, the sample sizes were too small (fewer than 20 patients for most laboratory data, particularly for CPK) to detect any significant variation.

Figure 5. Boxplots of laboratory results for the top 5 DDIs of each statin. The 3 control groups are depicted in purple, blue and green. Boxplots in yellow, orange and red indicate the laboratory results of patients exposed to statin-related DDI with a level of severity of 3, 2 and 1 (the most severe). Patients can have more than one DDI, and they can be of different severity. Fenofibrate and cyclosporin have 2 boxplots because some of their DDIs are classified as level 2 and others as level 1. ALT: alanine aminotransferase; AST: aspartate aminotransferase; CPK: creatine phosphokinase; DDI: drug–drug interaction.



Discussion

Strengths

Technical Work

To the best of our knowledge, this is the first study that successfully linked EHR data, through a CDW and claim data. However, there are some initiatives that integrate the 2 data types at the source into a common database [26,27].

The linkage process was efficient and generic enough to be applied to any data source that contains PMSI data. Our goal was to demonstrate that for data reuse purposes, it is possible to link fine-grained EHR data and claim data without a common patient identifier. Today, most hospitals have a CDW dedicated to research and fed with EHR data. Specifically, we used the eHOP CDW architecture that is currently the most widespread CDW type in France [13].

These 2 data sources can be bulky. For instance, the statin use-cases required to read and filter all drug administrations (n=13,125,574) and all drug dispensations (n=6,019,432) to identify patients to be included in the study were large. To ensure fast computation, we developed a computing framework based on Spark and HDFS that showed good performances even on our small single node cluster. These tools are widespread in the big data field, but they are still rarely used for data reuse in hospitals. According to Dolezel et al [28], their underuse, despite

the massive amount of hospital data available, is explained by the lack of personnel with specific technological skills.

DDI Use Case

Our use-case study found a statin-drug interaction prevalence of 22.17% (3253/14,675) and 36.45% (3800/10,424), during community consumption and hospital consumption, respectively. Few studies have provided statin-drug interaction rates during primary care and hospital care for the same population. A Bulgarian study [29] reported statin-drug interaction prevalence rates of 26.1% at hospital admission (used as a proxy for primary care prescription) and 24.4% at discharge. Regarding primary care, this rate ranges from 6.9% [30] to 33% in a systematic review [31] on elderly patients. However, the definition of interaction varies among studies. This could be explained not only by the choice of drug database, as reported in the literature [32,33], but also by the focus on the most severe interactions. Our study took into account different severity levels, from precaution of use to contraindication, using the Thériaque database.

By comparing the places where interactions occurred (community or hospital), our study showed that the most severe interactions in the hospital led to more specialized and longer care, as previously reported [34]. This should be put in perspective with the larger number and types of drugs administered during hospital stays. Finally, we attempted to link DDIs and laboratory results and showed their potential impact on some laboratory parameters. Previous works reported the biological effects of some statin-drug interactions, such as (1) liver toxicity (elevated alanine aminotransferase or aspartate aminotransferase) by interaction with cyclosporin, (2) hyperkalemia [34] with itraconazole or erythromycin [35], and (3) hyperglycemia with fusidic acid [36]. These findings should be interpreted with caution because some of them could be because of the adverse effects of statins [37] or of the other drug, such as itraconazole.

Limitations

Technical Work

The pairing procedure showed that the data life cycle introduced quality defects that explained the incomplete record linkage. We are still investigating the reasons for the match failures and how to explain quality data defects. The record linkage procedure could be improved using more sophisticated linkage strategies, such as probabilistic methods. However, our study concerned a specific case where data variables used for the record linkage procedure originated from the same source (ie, PMSI data produced by hospitals). Most of the unmatched patients were twins who could not be distinguished in the SNDS data, even by using more complex methods. We think that the deterministic approach is simpler to maintain and is more understandable for people who would like to use or adapt our algorithm for their own purpose.

DDI Use Case

DDI prevalence remains dependent on the chosen definition. In our study, these interactions were based only on the simple presence of a drug that could interact with statins and did not capture dose-dependency or patient-specific factors that might influence DDI definitions. Moreover, only information on dispensation was available for primary care (community consumption), whereas administrations were considered for hospital stay.

Despite the large cohort of patients over a 3-year period, our use case study found only 121 patients with a severity level 1 DDI, and among them only 5 had CPK data. This highlights the importance of the large sample size needed in pharmacoepidemiology and pharmacovigilance studies to detect rare adverse effects.

Conclusions

This study demonstrates the added value of combining and reusing clinical and claim data to provide large-scale measures of DDI prevalence and care pathways outside hospitals. In a complex health care system that involves multiple care providers, transitions of care are often the source of medication discrepancies and DDIs [38]. Linking CDW and community data is a promising approach to identify gaps in the system.

Our approach also allows performing big data–driven analyses to generate new hypotheses. For instance, by linking laboratory data with DDIs, we demonstrated that our strategy allowed exploring potential biological variations associated with DDI exposure. However, because of the small patient samples with laboratory results and the exploratory design of the study, we did not want to infer any causal effect or clinical impact at this step. In this context, data reuse should be complementary to hypothesis-driven pharmacoepidemiological research, which is the appropriate way to confirm the plausibility of a given hypothesis generated using health data.

This builds the path to progress toward a Learning Health System, in which patient care is continuously improved using knowledge generated from research on real-world health data and clinical research [39].

Since the INSHARE project, we have extended this approach in the HUGOSHARE project in which we plan to analyze, using the Health Data Hub platform [40], the DDIs for a larger number of drug classes in a much bigger data set from SNDS and from the CDWs of 6 academic hospitals of the French western area. This may overcome the limitations of this study concerning the limited sample sizes for rare events with the aim to generate high quality hypotheses and to consider building predictive models.

Future medical technological developments may also consider enriching community pharmacy reimbursement data with other community data, such as community laboratory results or ambulatory visits. This might enable researchers to identify system vulnerabilities that result in medication errors slipping through the holes of the *Swiss Cheese Model of System Errors* [41,42].

Acknowledgments

We would like to thank Daenna Wung, who contributed to this work.

This work was supported by the French National Research Agency for the INSHARE (Integrating and Sharing Health Big Data for Research) project (grant no. ANR-15-CE19-0024), and by Epiphare (Groupement d'intérêt scientifique) for the record linkage procedure between the Système National des Données de Santé and Hospital clinical data.

Authors' Contributions

MC, GB, and AB conceived of and designed the study. AB, GB, PL and PVH collected and analyzed the data. MC, GB, AB, MB, PLC, EC, XD, and CR interpreted the data. AB and GB drafted the manuscript, and MB, PLC, CR, PL, PVH, EC, XD, and MC critically revised the manuscript. MC obtained funding and supervised the study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Pseudocode describing the different steps of the record-linkage algorithm.
[DOCX File , 211 KB - medinform_v9i12e29286_app1.docx]

References

1. Chazard E, Ficheur G, Caron A, Lamer A, Labreuche J, Cuggia M, et al. Secondary use of healthcare structured data: the challenge of domain-knowledge based extraction of features. Stud Health Technol Inform 2018;255:15-19. [Medline: 30306898]
2. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform 2017 Aug;26(1):38-52 [FREE Full text] [doi: 10.15265/IY-2017-007] [Medline: 28480475]
3. Safran C. Reuse of clinical data. Yearb Med Inform 2014 Aug 15;9(1):52-54 [FREE Full text] [doi: 10.15265/IY-2014-0013] [Medline: 25123722]
4. Pacheco AG, Saraceni V, Tuboi SH, Moulton LH, Chaisson RE, Cavalcante SC, et al. Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil. Am J Epidemiol 2008 Dec 01;168(11):1326-1332 [FREE Full text] [doi: 10.1093/aje/kwn249] [Medline: 18849301]
5. Asher J, Resnick D, Brite J, Brackbill R, Cone J. An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. Int J Environ Res Public Health 2020 Sep 22;17(18):6937 [FREE Full text] [doi: 10.3390/ijerph17186937] [Medline: 32972036]
6. Sayers A, Ben-Shlomo Y, Blom AW, Steele F. Probabilistic record linkage. Int J Epidemiol 2016 Jun;45(3):954-964 [FREE Full text] [doi: 10.1093/ije/dyv322] [Medline: 26686842]
7. Oliveira GP, Bierrenbach AL, Camargo KR, Coeli CM, Pinheiro RS. Accuracy of probabilistic and deterministic record linkage: the case of tuberculosis. Rev Saude Publica 2016 Aug 22;50:49 [FREE Full text] [doi: 10.1590/S1518-8787.2016050006327] [Medline: 27556963]
8. Oostema JA, Nickles A, Reeves MJ. A comparison of probabilistic and deterministic match strategies for linking prehospital and in-hospital stroke registry data. J Stroke Cerebrovasc Dis 2020 Oct;29(10):105151. [doi: 10.1016/j.jstrokecerebrovasdis.2020.105151] [Medline: 32912531]
9. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 2011 May;64(5):565-572. [doi: 10.1016/j.jclinepi.2010.05.008] [Medline: 20952162]
10. Hagger-Johnson G, Harron K, Goldstein H, Aldridge R, Gilbert R. Probabilistic linkage to enhance deterministic algorithms and reduce data linkage errors in hospital administrative data. J Innov Health Inform 2017 Jun 30;24(2):891 [FREE Full text] [doi: 10.14236/jhi.v24i2.891] [Medline: 28749318]
11. Bosh KA, Coyle JR, Muriithi NW, Ramaswamy C, Zhou W, Brantley AD, et al. Linking HIV and viral hepatitis surveillance data: evaluating a standard, deterministic matching algorithm using data from 6 US health jurisdictions. Am J Epidemiol 2018 Nov 01;187(11):2415-2422. [doi: 10.1093/aje/kwy161] [Medline: 30099475]
12. Tuppin P, Rudant J, Constantinou P, Gastaldi-Ménager C, Rachas A, de Roquefeuil L, et al. Value of a national administrative database to guide public decisions: from the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. Rev Epidemiol Sante Publique 2017 Oct;65 Suppl 4:S149-S167. [doi: 10.1016/j.respe.2017.05.004] [Medline: 28756037]

13. Madec J, Bouzillé G, Riou C, Van Hille P, Merour C, Artigny M, et al. eHOP clinical data warehouse: from a prototype to the creation of an inter-regional clinical data centers network. Stud Health Technol Inform 2019 Aug 21;264:1536-1537. [doi: 10.3233/SHTI190522] [Medline: 31438219]
14. LOI n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles. Legifrance. 2020. URL: https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000036195293/ [accessed 2021-10-07]
15. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3:160018 [FREE Full text] [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
16. Ferrières J, Bongard V, Dallongeville J, Arveiler D, Cottel D, Haas B, et al. Trends in plasma lipids, lipoproteins and dyslipidaemias in French adults, 1996-2007. Arch Cardiovasc Dis 2009 Apr;102(4):293-301 [FREE Full text] [doi: 10.1016/j.acvd.2009.02.002] [Medline: 19427606]
17. Efficacité et efficience des hypolipémiants : une analyse centrée sur les statines. 2010. URL: https://www.has-sante.fr/jcms/r_1499450/fr/efficacite-et-efficience-des-hypolipemiants-une-analyse-centree-sur-les-statines [accessed 2021-10-07]
18. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, ESC Scientific Document Group. 2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk. Eur Heart J 2020 Jan 01;41(1):111-188. [doi: 10.1093/eurheartj/ehz455] [Medline: 31504418]
19. Stock J. Targeting LDL cholesterol: early treatment is key to population health. Atherosclerosis 2020 May;300:37-38. [doi: 10.1016/j.atherosclerosis.2020.03.013] [Medline: 32253009]
20. Law M, Rudnicka AR. Statin safety: a systematic review. Am J Cardiol 2006 Apr 17;97(8A):52C-60C. [doi: 10.1016/j.amjcard.2005.12.010] [Medline: 16581329]
21. Säfström E, Jaarsma T, Strömberg A. Continuity and utilization of health and community care in elderly patients with heart failure before and after hospitalization. BMC Geriatr 2018 Aug 13;18(1):177 [FREE Full text] [doi: 10.1186/s12877-018-0861-9] [Medline: 30103688]
22. Wastesson JW, Morin L, Tan EC, Johnell K. An update on the clinical consequences of polypharmacy in older adults: a narrative review. Expert Opin Drug Saf 2018 Dec;17(12):1185-1196. [doi: 10.1080/14740338.2018.1546841] [Medline: 30540223]
23. Leelakanok N, Holcombe AL, Lund BC, Gu X, Schweizer ML. Association between polypharmacy and death: a systematic review and meta-analysis. J Am Pharm Assoc (2003) 2017;57(6):729-38.e10. [doi: 10.1016/j.japh.2017.06.002] [Medline: 28784299]
24. Husson M. [Theriaque: independent-drug database for good use of drugs by health practitioners]. Ann Pharm Fr 2008;66(5-6):268-277. [doi: 10.1016/j.pharma.2008.07.009] [Medline: 19061726]
25. Bezin J, Duong M, Lassalle R, Droz C, Pariente A, Blin P, et al. The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf 2017 Aug;26(8):954-962. [doi: 10.1002/pds.4233] [Medline: 28544284]
26. Olatosi B, Zhang J, Weissman S, Hu J, Haider MR, Li X. Using big data analytics to improve HIV medical care utilisation in South Carolina: a study protocol. BMJ Open 2019 Jul 19;9(7):e027688 [FREE Full text] [doi: 10.1136/bmjopen-2018-027688] [Medline: 31326931]
27. McCoy RG, Dykhoff HJ, Sangaralingham L, Ross JS, Karaca-Mandic P, Montori VM, et al. Adoption of new glucose-lowering medications in the U.S.-the case of SGLT2 inhibitors: nationwide cohort study. Diabetes Technol Ther 2019 Dec;21(12):702-712 [FREE Full text] [doi: 10.1089/dia.2019.0213] [Medline: 31418588]
28. Dolezel D, McLeod A. Big data analytics in healthcare: investigating the diffusion of innovation. Perspect Health Inf Manag 2019 Jul 1;16(Summer):1a [FREE Full text] [Medline: 31423120]
29. Zhelyazkova-Savova M, Gancheva S, Sirakova V. Potential statin-drug interactions: prevalence and clinical significance. Springerplus 2014 Mar 31;3:168 [FREE Full text] [doi: 10.1186/2193-1801-3-168] [Medline: 24790817]
30. Rätz Bravo AE, Tchambaz L, Krähenbühl-Melcher A, Hess L, Schlienger RG, Krähenbühl S. Prevalence of potentially severe drug-drug interactions in ambulatory patients with dyslipidaemia receiving HMG-CoA reductase inhibitor therapy. Drug Saf 2005;28(3):263-275. [doi: 10.2165/00002018-200528030-00007] [Medline: 15733030]
31. Thai M, Reeve E, Hilmer SN, Qi K, Pearson S, Gnjidic D. Prevalence of statin-drug interactions in older people: a systematic review. Eur J Clin Pharmacol 2016 May;72(5):513-521. [doi: 10.1007/s00228-016-2011-7] [Medline: 26790666]
32. Fung KW, Kapusnik-Uner J, Cunningham J, Higby-Baker S, Bodenreider O. Comparison of three commercial knowledge bases for detection of drug-drug interactions in clinical decision support. J Am Med Inform Assoc 2017 Jul 01;24(4):806-812 [FREE Full text] [doi: 10.1093/jamia/ocx010] [Medline: 28339701]
33. Morival C, Westerlynck R, Bouzillé G, Cuggia M, Le Corre P. Prevalence and nature of statin drug-drug interactions in a university hospital by electronic health record mining. Eur J Clin Pharmacol 2018 Apr;74(4):525-534. [doi: 10.1007/s00228-017-2400-6] [Medline: 29255993]
34. Wang J, Chi C, St Peter WL, Carlson A, Loth M, Pradhan PM, et al. A population-based study of simvastatin drug-drug interactions in cardiovascular disease patients. AMIA Jt Summits Transl Sci Proc 2020;2020:664-673 [FREE Full text] [Medline: 32477689]

35. Li DQ, Kim R, McArthur E, Fleet JL, Bailey DG, Juurlink D, et al. Risk of adverse events among older adults following co-prescription of clarithromycin and statins not metabolized by cytochrome P450 3A4. CMAJ 2015 Feb 17;187(3):174-180 [FREE Full text] [doi: 10.1503/cmaj.140950] [Medline: 25534598]
36. Eng H, Scialis RJ, Rotter CJ, Lin J, Lazzaro S, Varma MV, et al. The antimicrobial agent fusidic acid inhibits organic anion transporting polypeptide-mediated hepatic clearance and may potentiate statin-induced myopathy. Drug Metab Dispos 2016 May;44(5):692-699. [doi: 10.1124/dmd.115.067447] [Medline: 26888941]
37. Kim J, Lee HS, Lee K. Effect of statins on fasting glucose in non-diabetic individuals: nationwide population-based health examination in Korea. Cardiovasc Diabetol 2018 Dec 05;17(1):155 [FREE Full text] [doi: 10.1186/s12933-018-0799-4] [Medline: 30518364]
38. Colombo F, Nunnari P, Ceccarelli G, Romano AV, Barbieri P, Scaglione F. Measures of drug prescribing at care transitions in an internal medicine unit. J Clin Pharmacol 2018 Sep;58(9):1171-1183. [doi: 10.1002/jcph.1123] [Medline: 29723431]
39. Platt JE, Raj M, Wienroth M. An analysis of the learning health system in its first decade in practice: scoping review. J Med Internet Res 2020 Mar 19;22(3):e17026 [FREE Full text] [doi: 10.2196/17026] [Medline: 32191214]
40. Cuggia M, Combes S. The French health data hub and the german medical informatics initiatives: two national projects to promote data sharing in healthcare. Yearb Med Inform 2019 Aug;28(1):195-202 [FREE Full text] [doi: 10.1055/s-0039-1677917] [Medline: 31419832]
41. Yao B, Kang H, Wang J, Zhou S, Gong Y. Toward reporting support and quality assessment for learning from reporting: a necessary data elements model for narrative medication error reports. AMIA Annu Symp Proc 2018 Dec 5;2018:1581-1590 [FREE Full text] [Medline: 30815204]
42. Wiegmann DA, J Wood L, N Cohen T, Shappell SA. Understanding the "Swiss Cheese Model" and its application to patient safety. J Patient Saf 2021 Apr 14 (forthcoming). [doi: 10.1097/PTS.0000000000000810] [Medline: 33852542]

Abbreviations

CDW: clinical data warehouse
CPK: creatine phosphokinase
DDI: drug–drug interaction
eHOP: entrepôt Hôpital
EHR: electronic health record
FAIR: findability, accessibility, interoperability, and reusability
HDFS: Hadoop Distributed File System
ICD-10: International Classification of Diseases, Tenth Revision
INSHARE: Integrating and Sharing Health Big Data for Research
PMSI: Programme de Médicalisation des Systèmes d'Information
SNDS: Système National des Données de Santé

Edited by C Lovis; submitted 31.03.21; peer-reviewed by R Tsopra, A Lamer; comments to author 02.05.21; revised version received 12.07.21; accepted 25.07.21; published 13.12.21.

Please cite as:
Bannay A, Bories M, Le Corre P, Riou C, Lemordant P, Van Hille P, Chazard E, Dode X, Cuggia M, Bouzillé G
Leveraging National Claims and Hospital Big Data: Cohort Study on a Statin-Drug Interaction Use Case
JMIR Med Inform 2021;9(12):e29286
URL: https://medinform.jmir.org/2021/12/e29286
doi:10.2196/29286
PMID:34898457

Original Paper

Machine Learning Methodologies for Prediction of Rhythm-Control Strategy in Patients Diagnosed With Atrial Fibrillation: Observational, Retrospective, Case-Control Study

Rachel S Kim[1], BA; Steven Simon[2], MD; Brett Powers[1], MSc; Amneet Sandhu[3], MD; Jose Sanchez[3], MD; Ryan T Borne[3], MD; Alexis Tumolo[3], MD; Matthew Zipse[3], MD; J Jason West[3], MD; Ryan Aleong[3], MD; Wendy Tzou[3], MD; Michael A Rosenberg[1,3], MD

[1]Colorado Center for Personalized Medicine, University of Colorado School of Medicine, Aurora, CO, United States

[2]Division of Cardiology, University of Colorado School of Medicine, Aurora, CO, United States

[3]Clinical Cardiac Electrophysiology Section, Division of Cardiology, University of Colorado School of Medicine, Aurora, CO, United States

Corresponding Author:
Michael A Rosenberg, MD
Clinical Cardiac Electrophysiology Section
Division of Cardiology
University of Colorado School of Medicine
12631 East 17th Avenue
Mail Stop B130
Aurora, CO, 80045
United States
Phone: 1 (303) 724 8391
Email: michael.a.rosenberg@cuanschutz.edu

Abstract

Background: The identification of an appropriate rhythm management strategy for patients diagnosed with atrial fibrillation (AF) remains a major challenge for providers. Although clinical trials have identified subgroups of patients in whom a rate- or rhythm-control strategy might be indicated to improve outcomes, the wide range of presentations and risk factors among patients presenting with AF makes such approaches challenging. The strength of electronic health records is the ability to build in logic to guide management decisions, such that the system can automatically identify patients in whom a rhythm-control strategy is more likely and can promote efficient referrals to specialists. However, like any clinical decision support tool, there is a balance between interpretability and accurate prediction.

Objective: This study aims to create an electronic health record–based prediction tool to guide patient referral to specialists for rhythm-control management by comparing different machine learning algorithms.

Methods: We compared machine learning models of increasing complexity and used up to 50,845 variables to predict the rhythm-control strategy in 42,022 patients within the University of Colorado Health system at the time of AF diagnosis. Models were evaluated on the basis of their classification accuracy, defined by the F1 score and other metrics, and interpretability, captured by inspection of the relative importance of each predictor.

Results: We found that age was by far the strongest single predictor of a rhythm-control strategy but that greater accuracy could be achieved with more complex models incorporating neural networks and more predictors for each participant. We determined that the impact of better prediction models was notable primarily in the rate of inappropriate referrals for rhythm-control, in which more complex models provided an average of 20% fewer inappropriate referrals than simpler, more interpretable models.

Conclusions: We conclude that any health care system seeking to incorporate algorithms to guide rhythm management for patients with AF will need to address this trade-off between prediction accuracy and model interpretability.

(JMIR Med Inform 2021;9(12):e29225) doi:10.2196/29225

KEYWORDS

atrial fibrillation; rhythm-control; machine learning; ablation; antiarrhythmia agents; data science; biostatistics; artificial intelligence

Introduction

Atrial Fibrillation

Atrial fibrillation (AF) affects an estimated 2.3 million Americans, with projections to over 10 million by the year 2050 [1,2], at current estimated costs of over US $26 billion each year in total [3] or US $18,000-US $20,000 per patient [4]. According to an analysis of the MarketScan database, patients diagnosed with AF underwent a mean 11.25 (SD 7.51) outpatient office visits, mean 4.74 (SD 5.24) outpatient hospital visits, and mean 0.71 (SD 1.28) emergency department visits, and were hospitalized for a mean 1.59 (SD 3.39) days on average over a given 6-month period [5]. Although the only treatment that has consistently reduced mortality from AF is the use of oral anticoagulation agents to prevent thromboembolic stroke [6-19], patients with AF can still have acute coronary syndromes, heart failure, and cardiovascular death at a rate of approximately 5% per year [20-23], including 35%-50% with hospital admissions or death within 5 years, even in the presence of oral anticoagulation [24,25]. Furthermore, the use of anticoagulation has no direct impact on the symptoms a patient may experience from AF, on the effect AF may have on underlying cardiovascular physiology, or on the long-term outcomes of being in AF rather than sinus rhythm. As such, the treatment of AF beyond identification of individuals needing anticoagulation is generally directed toward one of two strategies: (1) a rate-control strategy, focused solely on reducing the rate of ventricular excitation without attempting to restore sinus rhythm, or (2) a rhythm-control strategy, in which the focus is on restoring sinus rhythm using direct electrical energy (cardioversion), antiarrhythmic medications [26,27], catheter ablation, or a combination of two or more of these approaches [6,26,28-30]. Although a rate-control strategy can typically be performed under the care of a primary care physician, application of a rhythm-control strategy generally requires input from a specialist in cardiology or cardiac electrophysiology. Given the complexity of the decision about when to pursue a rhythm- or rate-control strategy, patients in whom a rhythm-control strategy is unlikely may be reflexively referred to cardiology or cardiac electrophysiology; in contrast, patients in whom a rhythm strategy would be beneficial may not be referred to a specialist who could provide this service. A method to identify patients who are more or less likely to undergo a rhythm-control strategy upstream could thus provide an attractive resource to improve care efficiency.

Use of Electronic Health Records

The expansion of electronic health records (EHRs) has created the opportunity to develop automated methods of prediction using machine learning. Although machine learning methods can provide superior predictability over standard methods in some cases, this improved accuracy often comes at the expense of using *black box* methods for prediction, in which it is not clear what specific information is being used by a given model to make predictions [31]. Within the space of clinical decision-making, such opacity can be a problem as it not only prevents users from gaining trust in the model but also provides little feedback in terms of how potential factors might be modified to change a decision. Our group has previously described the application of machine learning methods to EHRs for the prediction of incident AF and other outcomes [32,33].

In this study, we applied a step-by-step process to develop prediction models of increasing complexity using EHR data to predict whether a given patient is likely to have a rate- or rhythm-control strategy at the time of diagnosis of AF. We structured our analysis to examine and compare methods that offer a range of levels of model interpretability as well as prediction accuracy. In conclusion, we have provided a set of models that can be applied using EHR data at the point of care to guide referrals for AF management broadly within a health care system.

Methods

Study Population

The University of Colorado (UC) Health hospital system includes 3 large regional centers (north, central, and south) over the front range of Colorado. All UC Health hospitals share a single Epic instance, with backups and storage within Epic's Cogito Suite of databases, including Chronicles (operational database), Clarity (relational database), and Caboodle (dimensional database). In 2016, the UC entered into a unique partnership with Google to allow data from Caboodle to be loaded and stored in a research-focused data warehouse called the Health Data Compass, located entirely on the Google Cloud Platform, which was used by our team for this study. The data set was obtained using Google Big Query applied to the EHR system to return patients who were seen for outpatient encounters between October 11, 2010, and October 26, 2020, and were between the age group of 18 and 100 years at the index encounter, defined as the first time that a diagnosis of AF was entered for an outpatient seen at a UC Health clinic (see Multimedia Appendix 1, Table S1, for AF diagnosis definitions). The full data set contained 42,022 participants and was split into a training set (31,517/42,022, 75%) and a testing set (10,505/42,022, 25%), with model development performed using the training set and model comparisons using the testing set. This protocol was approved by the UC Multiple Institutional Review Board (#20-2192) using deidentified and uniquely encoded data sets, with a waiver of informed consent.

Clinical Predictors

Clinical predictors were grouped into two broad categories, which were defined as *big data predictors* and *known predictors*. *Big data predictors* included any diagnosis (International Classification of Disease [ICD]-9 or ICD-10) or procedure event for each patient before the index encounter, as well as race, ethnicity, and financial class. Any medication that was active and administered via the oral route at the index encounter was also included as a big data predictor. For each participant, an array was created for active medications, procedures, and diagnoses, followed by the use of a tokenizer (*Keras Tokenizer*) to create a one-hot encoded data set with each unique medication, procedure, and diagnosis assigned its own variable, resulting in a data set containing 50,845 variables. *Known predictors* were defined as any cardiac or metabolic diagnoses that have been identified as having a potential association with the risk of AF, including hypertension (ICD-9 401.X; ICD-10

I10.X) [6,21,34], obesity (ICD-9 278.X; ICD-10 E66.X) [34-37], diabetes mellitus (ICD-9 250.X; ICD-10 E11.X), coronary artery disease (ICD-9 414.X; ICD-10 I25.X), and heart failure (ICD-9 428.X; ICD-10 I50.X) [21,24,34,38,39], and mitral valve disease (ICD-9 424.X or 394.X, ICD-10 I34.X), as well as age and sex. Age was normalized (mean subtracted and divided by SD) for all analyses except for logistic regression models and decision trees (not including random forests [RFs]), which used the unnormalized age. This allowed for improved optimization of the models that used the normalized age and greater interpretability of the models that used the unnormalized age. Missing values were imputed using the median value (continuous variables) or mode (discrete variables). No participants were missing age or sex, and diagnoses were assumed to be absent if the value was unavailable.

Outcome: AF Treatment Strategy

AF treatments were defined as any medication, including antiarrhythmic medications, external cardioversion, or AF ablation procedure that was ordered within 6 months after the index encounter (Multimedia Appendix 1, Table S2). We defined the order for any antiarrhythmic medication, ablation, or cardioversion procedure as a *rhythm-control* strategy and any nodal agent or absence of a treatment order as a *rate-control* strategy. Treatments were only assessed following the index encounter (ie, the first outpatient visit at which the diagnosis of AF was entered); we did not examine subsequent treatments or study visits beyond the first 6 months after the index encounter. In one subanalysis, we examined the first selected rhythm-control strategy after the AF diagnosis, grouped into one of the following categories: antiarrhythmic medication, external cardioversion, and ablation.

Modeling Strategy

Model Development

As the total number of participants to whom a rhythm-control strategy was applied was relatively low (imbalanced data), we first compared four methods of resampling: synthetic minority oversampling technique (SMOTE) [40,41], random oversampling, random undersampling, and Tomek links undersampling [42], as well as the use of raw features. Resampling was performed only in the training set.

Model development proceeded from the most interpretable (logistic regression) to the most complex and opaque (combined methods incorporating neural networks in ensemble format). Originally, we planned to run all models on both groups of inputs, known and big data predictors. However, we found that only deep learning models provided predictive accuracy for big data predictors. Thus, we ran the non–deep learning models on the known predictors only (Multimedia Appendix 1, Table S7). For logistic regression, we used the training data set to develop binary logistic regression classifiers for models of rate- versus rhythm-control and multinomial logistic regression for models of the first AF treatment strategy among those identified as having a rhythm-control strategy. For RFs, extreme gradient boosting, K-nearest neighbors, and naïve Bayes classification, grid search for hyperparameter optimization was performed using five-fold cross-validation on the training set, with manual grid optimization to ensure that the grid contained the optimal hyperparameters (ie, if a hyperparameter value was identified on the upper end of the grid range, the grid was expanded to ensure that the overall optimal hyperparameter was not beyond the bounds of the grid space).

The approach to fitting neural networks was to first increase the complexity (lower learning rate and increased numbers of layers and neurons) to improve fit on the training data and then to include regularization methods (eg, decrease the learning rate and add dropout) as the out-of-sample loss began to increase, as noted in the examination of learning curves (Multimedia Appendix 1, Figure S1). We used feed-forward neural networks for deep learning architecture. Unless described otherwise, neural networks used fully connected layers with Elu activation (except the final layer, sigmoid), He initialization, L2 regularization (Penalty=0.01), dropout (20%), batch normalization, binary cross-entropy loss, Root Mean Square Propagation optimizer with learning rate=1e-4, ρ=0.9, and 50 training epochs with early stopping. Formal comparisons of predictive accuracy are presented; any model structure or hyperparameters that are not presented can be assumed to have provided inferior predictive accuracy compared with the presented models.

We also examined several ensemble methods by integrating the optimal model on the basis of big data predictors (from neural networks) with known predictors to allow interpretability of the impact of each component on the overall prediction accuracy. We first included the predicted probability of a rhythm-control strategy for each participant on the basis of the neural network as an input into either a RF or logistic regression, with SMOTE resampling for the training set. We also examined the weights and structure of the neural network with big data inputs combined with auxiliary input from known predictors concatenated at the final layer, followed by the addition of a fully connected layer (called *neural network combined*) with sigmoid output to predict rhythm-control strategy. Weights from pretrained layers of the former models were frozen, with training only on additional layers after the addition of known predictors.

Model Interpretation

Our main goal was to identify an optimal model to predict the probability of providers applying a rhythm-control strategy on the basis of classification accuracy and interpretability. Classification accuracy was defined primarily by the F1 score, with supportive metrics including the area under the receiver operator characteristic curve (AUC), precision (positive predictive value), recall (sensitivity), accuracy (% correct predictions), and inspection of the 2 × 2 contingency table. Interpretability was examined by inspecting the relative importance of each predictor according to the metrics available for each modeling approach. For logistic regression, importance was defined by the chi-square statistic from a nested likelihood ratio test, with and without inclusion of the predictor in the model. For RFs, importance was defined by the Gini index, which describes the mean decrease in impurity across all nodes, averaged over all decision trees [43]. We also examined individual decision trees manually for the interpretability and relevance of decision cut-points.

Model calibration was assessed using calibration curves created by binning the predicted probability from each model over the deciles of prediction and examining the actual proportion of rhythm-control strategies within each decile. Receiver operator characteristic and precision-recall curves were plotted using standard methods (*sklearn*). To allow inspection of these models within the context of triggering referrals for evaluation of the rhythm-control strategy, we also plotted the proportion of appropriate, inappropriate, and missed appropriate referrals according to varying probability thresholds from each prediction model. These classifications were assigned by comparing whether a rhythm-control strategy was predicted by the model and whether it was actually used for each participant. Thus, *appropriate referrals* indicated the participants for whom a rhythm-control strategy was predicted and used, *inappropriate referrals* indicated those for whom a rhythm-control strategy was predicted but not used (false positives), and *missed appropriate referrals* indicated those for whom a rhythm-control strategy was not predicted but was used (false negatives).

Computing Resources

Analyses and marginal estimation using logistic regression applied to the known predictors were conducted using Stata, IC (version 16, StataCorp, Inc). Analyses using both known and big data predictors were performed using scripts written in Python 3.7.4, with dependencies (software packages) including the following: *imblearn 0.0*, *Keras 2.2.4*, *numpy 1.19.4*, *pandas 0.25.1*, *scikit-learn 0.23.2*, and *tensorflow 2.4.0*. Scripts were developed and tested using Jupyter Notebook and deployed using command line programming at the UC's Health Data Compass Eureka virtual environment, hosted on Google Cloud Platform, using 64 central processing units and approximately 8-10 GB RAM, depending on the modeling requirements.

Results

Known Predictors

The overall study population demographics are provided in Table 1, split according to the strategy deployed (rate vs rhythm-control) and the training or testing set. A rhythm-control strategy was ordered within 6 months of AF diagnosis in 7.51% (3155/42,022) of patients. On average, patients undergoing a rhythm-control strategy were younger and male, with lower rates of existing cardiac conditions other than obesity. Among patients ordered for a rhythm-control strategy (and for whom this information was available), 20.88% (495/2370) were first ordered for ablation, 9.7% (230/2370) were ordered for an antiarrhythmic medication, and 69.41% (1645/2370) were ordered for external cardioversion. All known predictors (Table 1), except for obesity and hypertension, were significantly associated with a rhythm-control strategy at $P<.005$ (after Bonferroni adjustment for multiple comparisons). Nonlinearity of the interaction with age and sex was notable (Figure 1); younger men were more likely to have a rhythm-control strategy, with normalization of the sex-dependent effect by older age. Among the individuals in whom a rhythm-control strategy was ordered, the age-sex interaction remained significant, although the relationship between age and probability of rhythm-control strategy was no longer nonlinear (Figure 1). In addition, hypertension diagnosis was the strongest predictor of the type of rhythm-control strategy used. Individuals with a previous diagnosis of hypertension were less likely to have an ablation or antiarrhythmic medication and more likely to have a cardioversion ordered (Figure 1).

Table 1. Population demographics.

Demographics	Training set (n=31,517)		Testing set (n=10,505)	
	Rhythm control (n=2370)	Rate control (n=29,147)	Rhythm control (n=785)	Rate control (n=9720)
Age (years), mean (SD)	66.4 (12.0)	72.1 (12.9)	67.1 (11.6)	72.3 (12.7)
Sex (female), n (%)	779 (32.9)	12,588 (43.2)	265 (33.8)	4115 (42.3)
HTN[a], n (%)[b]	1036 (43.7)	14,577 (50)	372 (47.4)	4870 (50.1)
Obesity, n (%)[c]	366 (15.4)	3877 (13.3)	156 (19.9)	1243 (12.8)
Diabetes, n (%)[d]	343 (14.5)	5305 (18.2)	115 (14.7)	1768 (18.2)
CAD[e], n (%)[f]	475 (20)	7433 (24.5)	164 (20.9)	2497 (25.7)
Heart failure, n (%)[g]	488 (20.6)	5625 (19.3)	142 (18.1)	1874 (19.3)
Mitral valve disease, n (%)[h]	394 (16.6)	4841 (16.6)	124 (15.8)	1687 (17.4)

[a]HTN: hypertension diagnosis.

[b]International Classification of Disease-9 401.X; International Classification of Disease-10 I10.X.

[c]Obesity diagnosis (International Classification of Disease-9 278.X; International Classification of Disease-10 E66.X).

[d]Diabetes mellitus (International Classification of Disease-9 250.X; International Classification of Disease-10 E11.X).

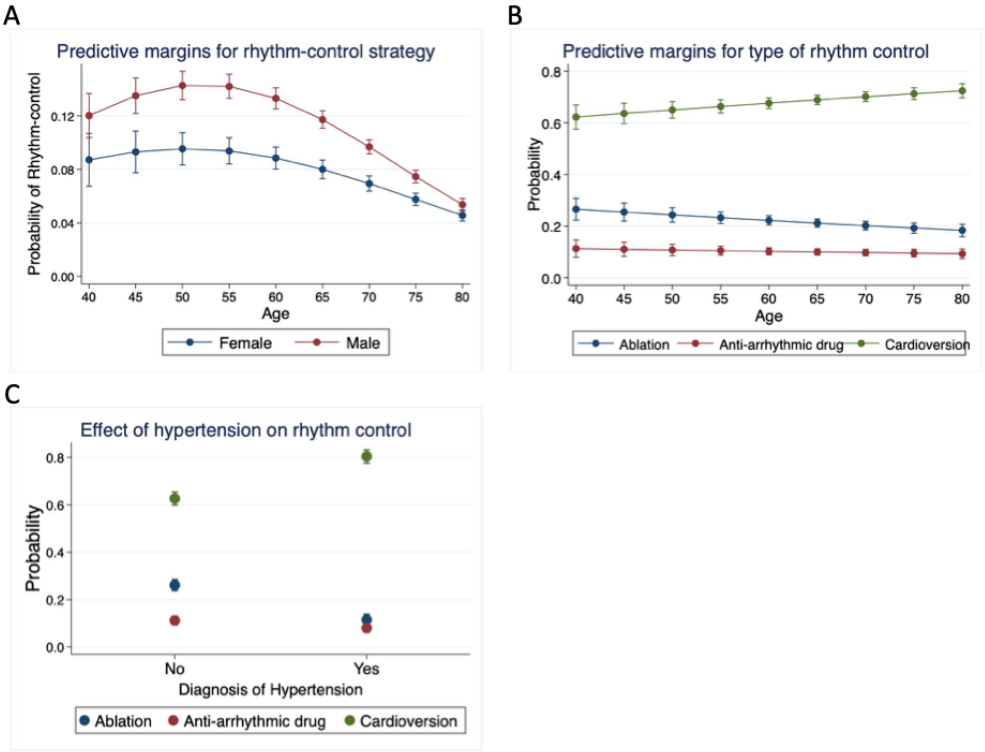
[e]CAD: coronary artery disease.

[f]International Classification of Disease-9 414.X; International Classification of Disease-10 I25.X.

[g]Heart failure (International Classification of Disease-9 428.X; International Classification of Disease-10 I50.X).

[h]Mitral valve disease (International Classification of Disease-9 424.X or 394.X; International Classification of Disease-10 I34.X).

Figure 1. (A) Predictive margins for rhythm-control strategy. Based on logistic regression with age and age-squared and age-sex interactions. Error bars represent the 95% CIs applied to each age-sex combination. (B) Predictive margins for the type of rhythm-control strategy: ablation, antiarrhythmic drug, and external cardioversion. Based on multinomial logistic regression for the first rhythm-control treatment applied, with age and age-squared and age-sex interactions. Error bars represent the 95% CI applied to each age-sex combination. (C) Predictive margins for the effect of hypertension diagnosis on the rhythm-control strategy. Based on multinomial logistic regression for the first rhythm-control treatment applied, with age and age-squared and age-sex interactions. Error bars represent the 95% CI applied to each age-sex combination.



Among the supervised learning algorithms to predict a rhythm-control strategy based only on known predictors (Multimedia Appendix 1, Table S3), we found that all methods had a similar magnitude of F1 score and that some resampling method (SMOTE being most common) was needed for optimal prediction (Table 2). Feature importance applied to the highest performing RF model demonstrated that age was by far the strongest predictor (Table 3). Inspection of the decision tree (Figure 2) indicated that age <70 years was strongly associated with a rhythm-control strategy, and age >89 years was strongly associated with the rate-control strategy. When the models were tested on age-stratified data, there was a slight improvement in the average AUC associated with increased age, but this was not statistically significant (Multimedia Appendix 1, Figure S3). The logistic regression results showed similar relative importance for the features, although RF favored coronary artery disease slightly more than sex as a predictor compared with the logistic regression, and mitral valve disease was relatively less important for regression than RF (Table 3).

Table 2. Best supervised learning models.

Model[a]	Resampling	F1 score	AUC[b]	Accuracy	Recall	Precision
Random forest[c]	SMOTE[d]	0.186	0.591	0.689	0.476	0.116
Extreme gradient boosting[e]	Random oversampling	0.179	0.591	0.614	0.563	0.106
K-nearest neighbors[f]	Random undersampling	0.181	0.605	0.541	0.682	0.105
Naïve Bayes[g]	SMOTE	0.184	0.602	0.596	0.609	0.108
Logistic regression	SMOTE	0.185	0.608	0.570	0.654	0.108

[a]All models except neural network applied to known predictors only.

[b]AUC: area under the receiver operator characteristic curve.

[c]Random forest hyperparameters: estimators=200, maximum features=8, maximum leaf nodes=300.

[d]SMOTE: synthetic minority oversampling technique.

[e]Extreme gradient boosting hyperparameters: booster=gbtree, η=0.9, γ=0, α=1, λ=0.

[f]K-nearest neighbors: N=500.

[g]Naïve Bayes: α=0.

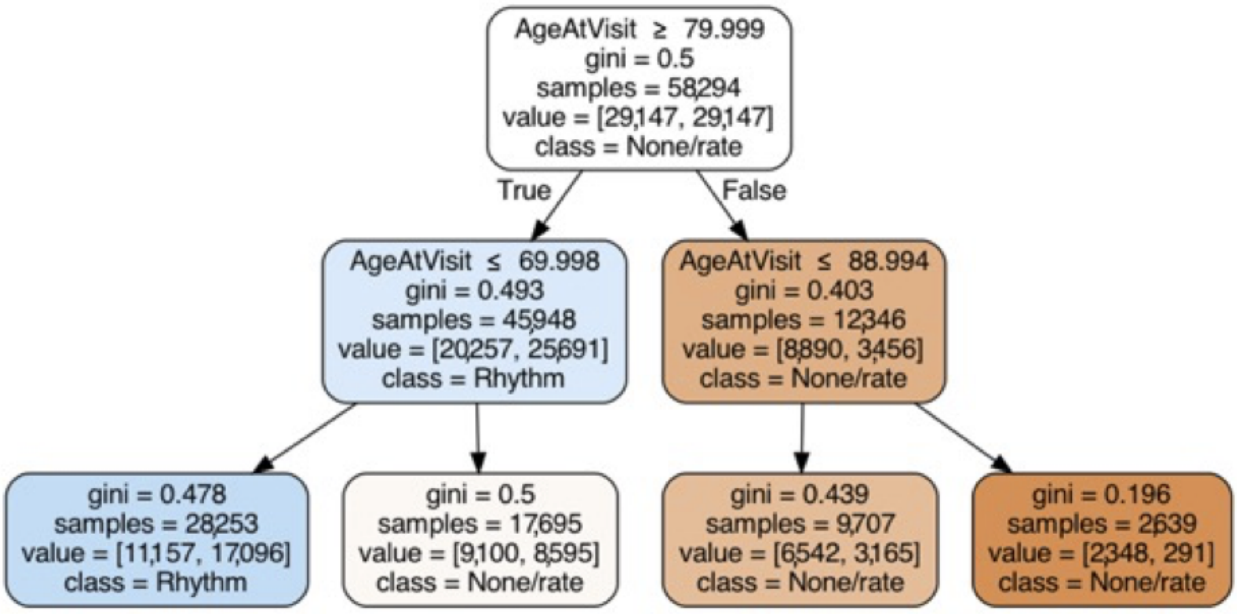
Table 3. Feature importance.

Predictor	Random forest impurity reduction[a] (%)	Logistic chi-square (*df*)	*P* value
Age (years)	81.74	462.11 (4)	<.001
CAD[b]	3.25	21.28 (1)	<.001
Sex	3.01	60.61 (3)	<.001
Mitral valve disease	2.82	8.04 (1)	.01
Diabetes mellitus	2.78	18.46 (1)	<.001
Heart failure	2.43	17.59 (1)	<.001
Hypertension	2.36	4.03 (1)	.04
Obesity	1.62	2.61 (1)	.11

[a]For random forest (synthetic minority oversampling technique resampling).

[b]CAD: coronary artery disease.

Figure 2. Decision tree for rhythm-control strategy. Based on known predictors to classify rate- versus rhythm-control strategy using the training data. Maximum depth=2, minimum samples to split nodes=50.



Big Data Predictors

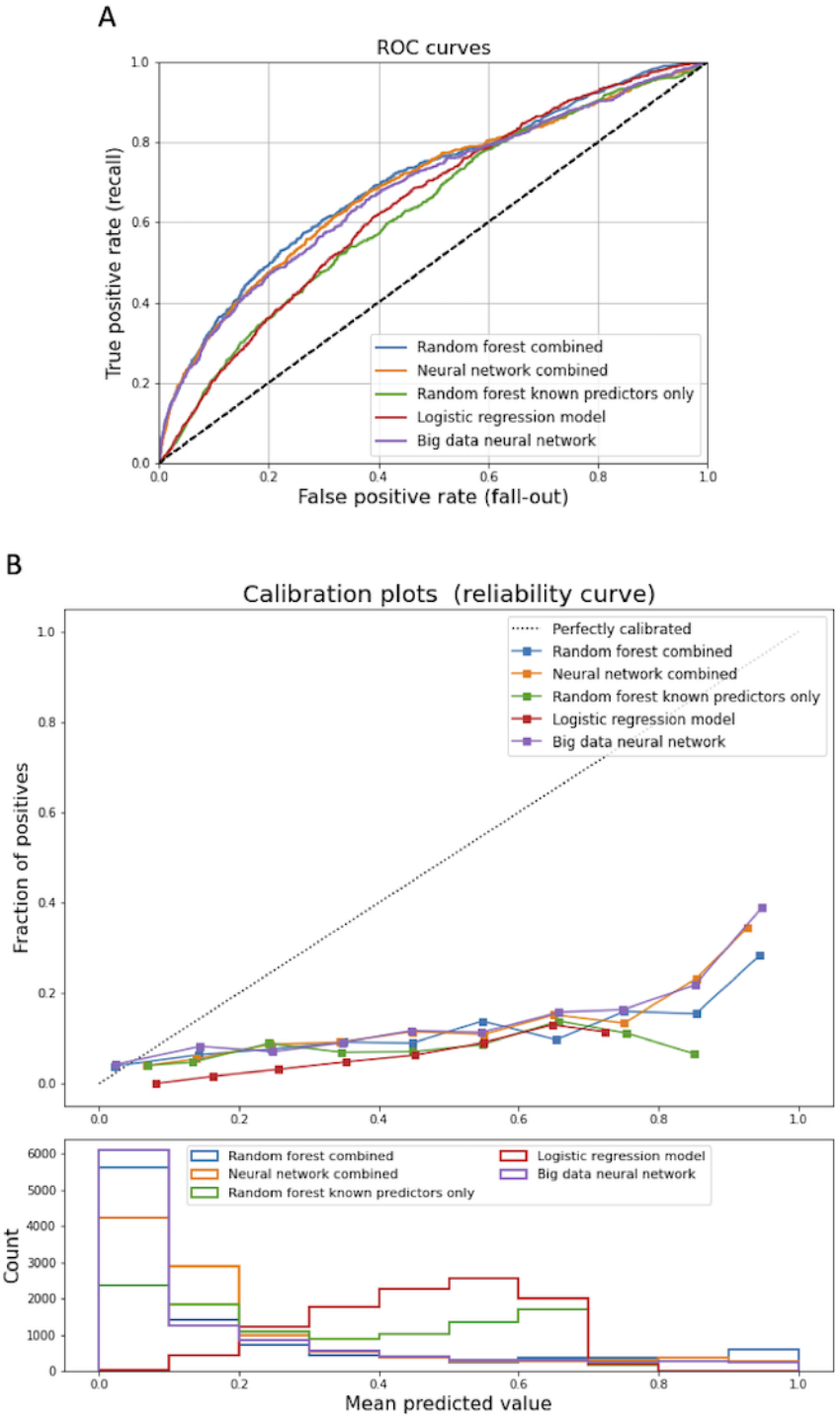
For big data predictors, only neural networks provided an F1 score over 0.0, so we focused on identifying the optimal neural network to predict a rhythm-control strategy. Across all neural networks using raw features, SMOTE, or random undersampling, we found that a 2-layer neural network with SMOTE provided superior prediction accuracy on the basis of the F1 score (Multimedia Appendix 1, Table S4). When examined within the context of logistic regression, decision tree, and RF, predictions from the big data neural network were by far the most predictive (Multimedia Appendix 1, Table S5). When combined as an ensemble with RF (RF combined) and neural network (neural network combined), the predictive accuracy remained high, with comparable F1 scores across models (Table 4) and clear improvement in prediction compared with RF or logistic regression based only on known predictors (Figure 3). Examination of calibration (Figure 3) indicated that all models were poorly calibrated and tended to overfit the data (predict rhythm-control strategy more often than this strategy was ordered).

Table 4. Combined big data (BD) and known predictor models.

Model	F1 score	AUC[a]	Accuracy	Recall	Precision
Random forests combined	0.258	0.643	0.807	0.451	0.181
Neural network combined	0.250	0.617	0.843	0.350	0.194
Neural network (BD predictors)	0.260	0.629	0.835	0.387	0.195

[a]AUC: area under the receiver operator characteristic curve.

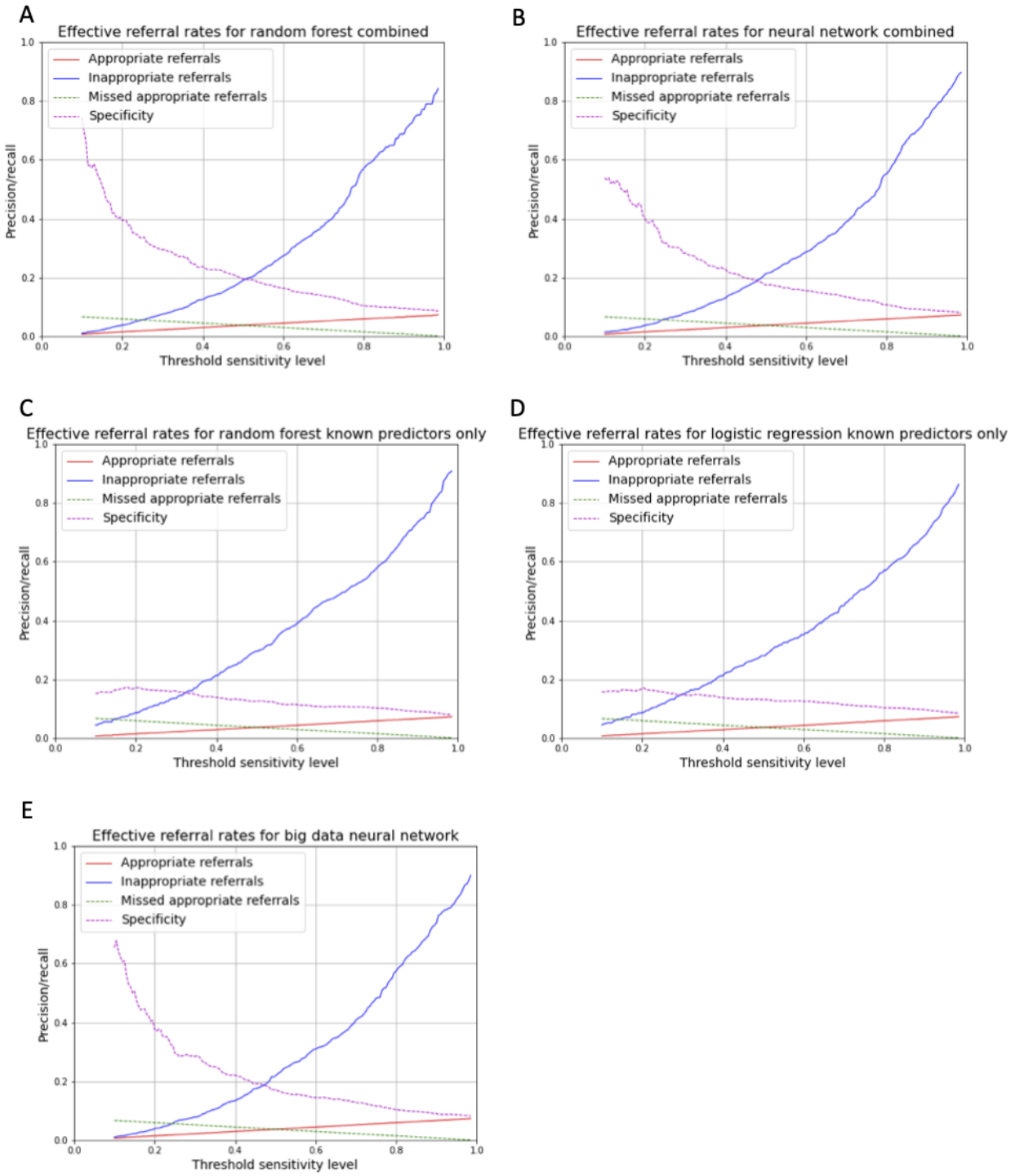
Figure 3. (A) Receiver operator characteristic curves for prediction models. Shown are top five models, including random forest combined and neural network combined (use big data and known inputs), random forest and logistic regression (use only known inputs), and neural network (only big data inputs). (B) Calibration curves (top) and histograms (bottom) for prediction models. Shown are top five models, including random forest combined and neural network combined (use big data and known inputs), random forest and logistic regression (use only known inputs), and neural network (only big data inputs). ROC: Receiver operator characteristic.



On the basis of precision-recall analysis (Multimedia Appendix 1, Figure S2), we examined the rate of appropriate, inappropriate, and missed appropriate referrals that would result from implementing an automated algorithm using these models at the time of AF diagnosis (Figure 4; Multimedia Appendix 1, Table S6). As expected, we found that the proportion of appropriate referrals (referral when rhythm-control strategy is likely) increased and missed appropriate referrals decreased with an increase in the sensitivity (recall) threshold used to guide the decision. However, it was also found that more complex models, such as those using combined known and big data predictors within a black box context, had a lower rate of

inappropriate referrals for thresholds between 0.3 and 0.8. To put this in context, if the model was applied to 10,000 patients at the time of AF diagnosis, increasing the sensitivity (recall) threshold from 0.5 to 0.7 would decrease the number of missed appropriate referrals by 150 patients for both models, at the expense of an increase in the number of inappropriate referrals of 1690 (logistic regression) to 1850 (RF combined). The use of models based solely on known predictors would increase the proportion of inappropriate referrals by approximately 20% compared with those that included big data predictors (Figure 4; Multimedia Appendix 1, Table S6).

Figure 4. Decision curves for prediction models based on proportion of appropriate and inappropriate referrals that would result from applying the model at different levels of sensitivity (thresholds): (A) random forest combined, (B) neural network combined, (C) random forest, (D) logistic regression, and (E) neural network.



Discussion

Principal Findings

In this EHR-based observational study of automated algorithms for the prediction of a rhythm-control strategy, we made several observations about the modeling process and the impact of using greater amounts of data to guide referrals. First, we found that nearly all methods were significantly improved by integration of some form of resampling during training (SMOTE being the most effective generally), which has been described previously by our group and others for the prediction of imbalanced outcomes. Although these approaches tended to improve the prediction accuracy as assessed by the F1 score and other measures of classification, they resulted in models that tended to predict a rhythm-control strategy more often than one was actually used, suggesting that they were overfitting the data. This result is consistent with previous work using machine learning to predict rare outcomes from EHR data by our team, including the prediction of AF itself [33] and myocardial infarction [32].

Second, we found that only neural networks could provide the computational power to produce accurate prediction models with big data inputs; none of the other approaches provided an AUC over 0.5 (F1 score>0.0) when applied to big data inputs. This result is also similar to previous findings with the application of machine learning to EHR data [32,33] and suggests the power of deep learning over standard methods, which has been demonstrated widely across a range of applications [44-46].

Finally, and most interestingly, we found that although no method was clearly superior to the others, there appeared to be a trade-off in which more interpretable models on the basis of known predictors alone provided inferior predictive accuracy compared with the use of more opaque, black box approaches incorporating deep neural networks. Specifically, we found that a model based solely on age could be reasonably effective for identifying patients in whom a rhythm-control strategy could be applied, but that greater levels of predictive accuracy required incorporation of much larger amounts of information, at the expense of not knowing which specific predictors (diagnoses, medications, or prior procedures) among the over 50,000 were needed. The benefit of using these more complex models was evident in a lower rate of inappropriate referrals within a wider range of thresholds, in which increasing the sensitivity of the predictions to decrease the number of missed appropriate referrals resulted in approximately 20% more inappropriate referrals for all but the lowest and highest thresholds. The bottom line is that a health system seeking to implement a clinical decision support algorithm could find a substantial increase in the costs due to inappropriate referrals in order to apply a more interpretable approach to guiding clinical decisions.

This study offers several comments, and the broader implications applied to both decisions about rhythm-control strategies and the role of machine learning and statistical modeling in EHR-based clinical decision support. In terms of rhythm- versus rate-control strategies, there are little data about the best approach for a given patient at the time of AF diagnosis. Early clinical trials limited to antiarrhythmic medications for rhythm-control showed no difference in outcomes for rhythm-control compared with a rate-control strategy [47-50], although more recent trials that include AF ablation for rhythm-control have noted improvements in ventricular function [21,38] and lower rates of stroke and death among patients with heart failure treated using a rhythm-control strategy that included AF ablation [24,51-53]. The recently published Early Treatment of Atrial Fibrillation for Stroke Prevention Trial 4 [54] examined early application of a rhythm-control strategy (within a year) and noted a reduction in the combined outcome of cardiovascular death, stroke, or cardiac hospitalizations [48,50,55], although the study did not directly measure costs [55]. Within the context of an automated referral algorithm, increasing the number of referrals blindly across the population is unlikely to be cost-effective, as we found that there was overall a relatively low rate (3155/42,022, 7.51%) of patients who had a rhythm-control strategy ordered within a 6-month period. In contrast, a program that avoids referrals for rhythm-control due to the overall low rate is likely to result in many patients being denied the opportunity to undergo treatment that could improve morbidity and mortality. We did not specifically examine long-term outcomes in this investigation, although we anticipate that like many models of automated decision-making, the procedure must start by mimicking expert decisions before moving on to models that incorporate outcomes. For example, the AlphaGo computer algorithm for playing Go began with modeling expert moves in the first version [56] before using automated game simulation to identify a model that could achieve suprahuman performance [57].

With regard to the use of deep learning models to make predictions about clinical decisions, there is an important issue of out-of-sample predictive accuracy, which includes model overfitting—fitting noise in the training data set that results in reduced predictive accuracy in the testing and validation data set—as well as sampling bias related to the population used to derive the prediction model being different from that in which it is applied. One of the remarkable features of modern deep learning methods is that through regularization techniques, such as dropout, these models are capable of fitting data in which the number of trainable parameters is greater than the number of samples or participants. However, due to the *curse of dimensionality*, the use of such a large number of predictors results in a large space of extrapolation (few data points *nearby* one another), in such a manner that only through trial and error, and use of strictly held-out testing data sets, can one increase the probability of fitting signal rather than noise. Even with careful attention to learning curves, one still cannot be certain of a model's predictive robustness without continued validation in external data sets. Such work is planned for these models, in which the trade-off between the use of a simple model with highly mappable inputs but lower predictive accuracy is balanced against the use of a complex deep learning model with greater accuracy; however, this requires a method to directly map approximately 50,000 features to the model input for application. Ultimately, more work will be needed to understand both the conceptual challenges of deep learning for clinical decision-making related to bias and overfitting, as well as the

practical issues of how one applies a model developed in one EHR to another.

Limitations

As expected from the examination of clinical decision-making using EHR data alone, there are several limitations to our study. First, as a result of the sheer number of encounters analyzed, we were unable to provide a manual chart or clinical validation of the decisions made in terms of rate or rhythm-control. As we defined the first diagnosis of AF as the first time it was entered into the EHR, it is highly likely that participants may have had undocumented AF before the index encounter and that a rate- or rhythm-control strategy may have been addressed at that point in time or by providers outside of our health care system. In addition, it is possible that many AF diagnoses were made in error and that patients may have had atrial flutter or supraventricular tachycardia rather than AF, in which case rate versus rhythm-control decisions would be irrelevant. Although we have an ongoing project to examine decisions at a patient-by-patient level, such an approach would not scale for the purposes of this analysis. Second, we selected an arbitrary 6-month window over which to assign a patient to a given strategy on the basis of whether a known rhythm-control approach was ordered. We were thus blind to patients who might have undergone a rhythm-control strategy outside the 6-month window or patients who started out with a rhythm-control strategy but then changed to rate-control going forward. Finally, although we were able to collect EHR-based data to apply predictive models, we were unable to obtain perhaps more relevant data pertaining to the decision about rate or rhythm control as it is applied clinically, such as symptoms or patterns of AF presentation. Clinically, symptoms are among the strongest reasons for referral for evaluation of AF by experts, and the inability to measure the symptoms with which a patient presents and how they progress is a limitation of our approach. Additional work using natural language processing of clinical notes or integration of other types of data related to patient activity or symptoms could provide a solution, although such data were not available at the time of this analysis. Importantly, the combined methodology we have described could be easily expanded to include this information without the need to retrain models entirely and could be directly analyzed in the same manner in which we integrated known predictors of AF alongside 50,000 big data inputs for prediction.

Conclusions

Historically, the direct application of clinical decision models was limited by data input capacity, integration of analytics with data storage, and the inability to deliver results directly at the point of care. However, advances in computer technology over the past 30 years have provided solutions to these problems toward the goal of incorporating artificial intelligence into clinical decision-making. The recent expansion of EHR use now provides vast amounts of data that can be collected, stored, and applied for clinical prediction at the point of care, without the need for manual data entry. These advances have created the opportunity for fully integrated artificial intelligence–based decision analysis at a scale previously unseen in clinical investigations, as well as allowing for dynamic updating of prediction models over time as greater amounts of data are collected and technologies and treatment options expand. This study is among the first to apply machine learning within the clinical decision context using this massive amount of data in a manner that could be directly applied within a health care system. The trade-off between model interpretability and predictive accuracy that we found is likely to be repeated across many future applications in which understanding the role of predictors is balanced against thousands, and potentially millions, of dollars in unnecessary referrals if such a system were automated. Clearly, more work is required before these systems can be implemented without oversight from a clinician; however, as we have noted, administrators and health care decision-makers should be aware that there is likely to arise a situation in which interpretability comes with a cost.

Acknowledgments

The authors would like to thank Rashawnda Franklin, Wenxin Wu, Michelle Edelmann, and Ian Brooks of the University of Colorado Health Data Compass team for providing the data used for this analysis. This work was supported by grants from the National Institutes of Health (R01 HL146824, K23 HL127296).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplemental figures and tables.
[PDF File (Adobe PDF File), 3652 KB - medinform_v9i12e29225_app1.pdf]

References

1. Miyasaka Y, Barnes M, Gersh B, Cha SS, Seward JB, Bailey KR, et al. Time trends of ischemic stroke incidence and mortality in patients diagnosed with first atrial fibrillation in 1980 to 2000. Stroke 2005 Oct 13;36(11):2362-2366. [doi: 10.1161/01.str.0000185927.63746.23]

2. Piccini JP, Hammill BG, Sinner MF, Jensen PN, Hernandez AF, Heckbert SR, et al. Incidence and prevalence of atrial fibrillation and associated mortality among Medicare beneficiaries, 1993-2007. Circ Cardiovasc Qual Outcomes 2012 Jan;5(1):85-93 [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.111.962688] [Medline: 22235070]
3. Kim MH, Johnston SS, Chu B, Dalal MR, Schulman KL. Estimation of total incremental health care costs in patients with atrial fibrillation in the United States. Circ Cardiovasc Qual Outcomes 2011 May;4(3):313-320. [doi: 10.1161/CIRCOUTCOMES.110.958165] [Medline: 21540439]
4. Delaney JA, Yin X, Fontes JD, Wallace ER, Skinner A, Wang N, et al. Hospital and clinical care costs associated with atrial fibrillation for Medicare beneficiaries in the Cardiovascular Health Study and the Framingham Heart Study. SAGE Open Med 2018 Feb 20;6:2050312118759444 [FREE Full text] [doi: 10.1177/2050312118759444] [Medline: 29511541]
5. Ladapo JA, David G, Gunnarsson CL, Hao SC, White SA, March JL, et al. Healthcare utilization and expenditures in patients with atrial fibrillation treated with catheter ablation. J Cardiovasc Electrophysiol 2012 Jan;23(1):1-8. [doi: 10.1111/j.1540-8167.2011.02130.x] [Medline: 21777324]
6. Packer DL, Mark DB, Robb RA, Monahan KH, Bahnson TD, Poole JE, CABANA Investigators. Effect of catheter ablation vs antiarrhythmic drug therapy on mortality, stroke, bleeding, and cardiac arrest among patients with atrial fibrillation: the CABANA randomized clinical trial. JAMA 2019 Apr 02;321(13):1261-1274 [FREE Full text] [doi: 10.1001/jama.2019.0693] [Medline: 30874766]
7. Boston Area Anticoagulation Trial for Atrial Fibrillation Investigators, Singer DE, Hughes RA, Gress DR, Sheehan MA, Oertel LB, et al. The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation. N Engl J Med 1990 Nov 29;323(22):1505-1511. [doi: 10.1056/NEJM199011293232201] [Medline: 2233931]
8. No author listed. Stroke prevention in atrial fibrillation study. Final results. Circulation 1991 Aug;84(2):527-539. [doi: 10.1161/01.cir.84.2.527] [Medline: 1860198]
9. No author listed. Warfarin versus aspirin for prevention of thromboembolism in atrial fibrillation: Stroke Prevention in Atrial Fibrillation II Study. Lancet 1994 Mar 19;343(8899):687-691. [Medline: 7907677]
10. Petersen P, Boysen G, Godtfredsen J, Andersen E, Andersen B. Placebo-controlled, randomised trial of warfarin and aspirin for prevention of thromboembolic complications in chronic atrial fibrillation. The Copenhagen AFASAK study. Lancet 1989 Jan 28;1(8631):175-179. [doi: 10.1016/s0140-6736(89)91200-2] [Medline: 2563096]
11. Ezekowitz MD, Bridgers SL, James KE, Carliner NH, Colling CL, Gornick CC, et al. Warfarin in the prevention of stroke associated with nonrheumatic atrial fibrillation. Veterans Affairs Stroke Prevention in Nonrheumatic Atrial Fibrillation Investigators. N Engl J Med 1992 Nov 12;327(20):1406-1412. [doi: 10.1056/NEJM199211123272002] [Medline: 1406859]
12. Connolly SJ, Laupacis A, Gent M, Roberts RS, Cairns JA, Joyner C. Canadian Atrial Fibrillation Anticoagulation (CAFA) study. J Am Coll Cardiol 1991 Aug;18(2):349-355 [FREE Full text] [doi: 10.1016/0735-1097(91)90585-w] [Medline: 1856403]
13. No author listed. Risk factors for stroke and efficacy of antithrombotic therapy in atrial fibrillation. Analysis of pooled data from five randomized controlled trials. Arch Intern Med 1994 Jul 11;154(13):1449-1457. [Medline: 8018000]
14. Hart RG, Pearce LA, Aguilar MI. Meta-analysis: antithrombotic therapy to prevent stroke in patients who have nonvalvular atrial fibrillation. Ann Intern Med 2007 Jun 19;146(12):857-867. [doi: 10.7326/0003-4819-146-12-200706190-00007] [Medline: 17577005]
15. van Walraven C, Hart RG, Singer DE, Laupacis A, Connolly S, Petersen P, et al. Oral anticoagulants vs aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. JAMA 2002 Nov 20;288(19):2441-2448. [doi: 10.1001/jama.288.19.2441] [Medline: 12435257]
16. Cooper NJ, Sutton AJ, Lu G, Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. Arch Intern Med 2006 Jun 26;166(12):1269-1275. [doi: 10.1001/archinte.166.12.1269] [Medline: 16801509]
17. Connolly SJ, Ezekowitz MD, Yusuf S, Eikelboom J, Oldgren J, Parekh A, RE-LY Steering CommitteeInvestigators. Dabigatran versus warfarin in patients with atrial fibrillation. N Engl J Med 2009 Sep 17;361(12):1139-1151. [doi: 10.1056/NEJMoa0905561] [Medline: 19717844]
18. Patel MR, Mahaffey KW, Garg J, Pan G, Singer DE, Hacke W, ROCKET AF Investigators. Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. N Engl J Med 2011 Sep 08;365(10):883-891. [doi: 10.1056/NEJMoa1009638] [Medline: 21830957]
19. Connolly SJ, Eikelboom J, Joyner C, Diener HC, Hart R, Golitsyn S, AVERROES Steering CommitteeInvestigators. Apixaban in patients with atrial fibrillation. N Engl J Med 2011 Mar 03;364(9):806-817. [doi: 10.1056/NEJMoa1007432] [Medline: 21309657]
20. Marijon E, Le Heuzey J, Connolly S, Yang S, Pogue J, Brueckmann M, RE-LY Investigators. Causes of death and influencing factors in patients with atrial fibrillation: a competing-risk analysis from the randomized evaluation of long-term anticoagulant therapy study. Circulation 2013 Nov 12;128(20):2192-2201. [doi: 10.1161/CIRCULATIONAHA.112.000491] [Medline: 24016454]
21. Willems S, Meyer C, de Bono J, Brandes A, Eckardt L, Elvan A, et al. Cabins, castles, and constant hearts: rhythm control therapy in patients with atrial fibrillation. Eur Heart J 2019 Dec 07;40(46):3793-379c [FREE Full text] [doi: 10.1093/eurheartj/ehz782] [Medline: 31755940]

22. Kirchhof P, Radaideh G, Kim Y, Lanas F, Haas S, Amarenco P, Global XANTUS program Investigators. Global prospective safety analysis of rivaroxaban. J Am Coll Cardiol 2018 Jul 10;72(2):141-153 [FREE Full text] [doi: 10.1016/j.jacc.2018.04.058] [Medline: 29976287]
23. Ruff CT, Giugliano RP, Braunwald E, Hoffman EB, Deenadayalu N, Ezekowitz MD, et al. Comparison of the efficacy and safety of new oral anticoagulants with warfarin in patients with atrial fibrillation: a meta-analysis of randomised trials. Lancet 2014 Mar 15;383(9921):955-962. [doi: 10.1016/S0140-6736(13)62343-0] [Medline: 24315724]
24. Packer D, Mark D, Robb R. Effect of Catheter Ablation vs Antiarrhythmic Drug Therapy on Mortality, Stroke, Bleeding, and Cardiac Arrest Among Patients With Atrial Fibrillation: The CABANA Randomized Clinical Trial. JAMA. Apr 2019;321(13):2-1274. [doi: 10.3410/f.735328400.793558685]
25. Hohnloser SH, Crijns HJ, van Eickels M, Gaudin C, Page RL, Torp-Pedersen C, ATHENA Investigators. Effect of dronedarone on cardiovascular events in atrial fibrillation. N Engl J Med 2009 Feb 12;360(7):668-678. [doi: 10.1056/NEJMoa0803778] [Medline: 19213680]
26. Calkins H, Reynolds MR, Spector P, Sondhi M, Xu Y, Martin A, et al. Treatment of atrial fibrillation with antiarrhythmic drugs or radiofrequency ablation: two systematic literature reviews and meta-analyses. Circ Arrhythm Electrophysiol 2009 Aug;2(4):349-361. [doi: 10.1161/CIRCEP.108.824789] [Medline: 19808490]
27. Lafuente-Lafuente C, Mouly S, Longás-Tejero MA, Mahé I, Bergmann J. Antiarrhythmic drugs for maintaining sinus rhythm after cardioversion of atrial fibrillation: a systematic review of randomized controlled trials. Arch Intern Med 2006 Apr 10;166(7):719-728. [doi: 10.1001/archinte.166.7.719] [Medline: 16606807]
28. Barnett AS, Kim S, Fonarow GC, Thomas LE, Reiffel JA, Allen LA, et al. Treatment of atrial fibrillation and concordance with the American Heart Association/American College of Cardiology/Heart Rhythm Society guidelines: findings from ORBIT-AF (Outcomes Registry for Better Informed Treatment of Atrial Fibrillation). Circ Arrhythm Electrophysiol 2017 Nov;10(11):e005051. [doi: 10.1161/CIRCEP.117.005051] [Medline: 29141842]
29. Jaïs P, Cauchemez B, Macle L, Daoud E, Khairy P, Subbiah R, et al. Catheter ablation versus antiarrhythmic drugs for atrial fibrillation: the A4 study. Circulation 2008 Dec 09;118(24):2498-2505. [doi: 10.1161/CIRCULATIONAHA.108.772582] [Medline: 19029470]
30. Oral H, Scharf C, Chugh A, Hall B, Cheung P, Good E, et al. Catheter ablation for paroxysmal atrial fibrillation: segmental pulmonary vein ostial ablation versus left atrial ablation. Circulation 2003 Nov 11;108(19):2355-2360. [doi: 10.1161/01.CIR.0000095796.45180.88] [Medline: 14557355]
31. Lipton ZC. The mythos of model interpretability. Queue 2018;16(3):31-57. [doi: 10.1145/3236386.3241340]
32. Mandair D, Tiwari P, Simon S, Colborn KL, Rosenberg MA. Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data. BMC Med Inform Decis Mak 2020 Oct 02;20(1):252 [FREE Full text] [doi: 10.1186/s12911-020-01268-x] [Medline: 33008368]
33. Tiwari P, Colborn KL, Smith DE, Xing F, Ghosh D, Rosenberg MA. Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation. JAMA Netw Open 2020 Jan 03;3(1):e1919396 [FREE Full text] [doi: 10.1001/jamanetworkopen.2019.19396] [Medline: 31951272]
34. Noheria A, Shrader P, Piccini JP, Fonarow GC, Kowey PR, Mahaffey KW, ORBIT-AF InvestigatorsPatients. Rhythm control versus rate control and clinical outcomes in patients with atrial fibrillation: results from the ORBIT-AF registry. JACC Clin Electrophysiol 2016 Apr;2(2):221-229 [FREE Full text] [doi: 10.1016/j.jacep.2015.11.001] [Medline: 29766874]
35. Ardestani A, Hoffman HJ, Cooper HA. Obesity and outcomes among patients with established atrial fibrillation. Am J Cardiol 2010 Aug 01;106(3):369-373 [FREE Full text] [doi: 10.1016/j.amjcard.2010.03.036] [Medline: 20643247]
36. Badheka AO, Rathod A, Kizilbash MA, Garg N, Mohamad T, Afonso L, et al. Influence of obesity on outcomes in atrial fibrillation: yet another obesity paradox. Am J Med 2010 Jul;123(7):646-651. [doi: 10.1016/j.amjmed.2009.11.026] [Medline: 20609687]
37. Packer M. Disease-treatment interactions in the management of patients with obesity and diabetes who have atrial fibrillation: the potential mediating influence of epicardial adipose tissue. Cardiovasc Diabetol 2019 Sep 24;18(1):121 [FREE Full text] [doi: 10.1186/s12933-019-0927-9] [Medline: 31551089]
38. Marrouche NF, Brachmann J, Andresen D, Siebels J, Boersma L, Jordaens L, CASTLE-AF Investigators. Catheter ablation for atrial fibrillation with heart failure. N Engl J Med 2018 Feb 01;378(5):417-427. [doi: 10.1056/NEJMoa1707855] [Medline: 29385358]
39. Choi YJ, Kang K, Kim T, Cha M, Lee J, Park J, et al. Comparison of rhythm and rate control strategies for stroke occurrence in a prospective cohort of atrial fibrillation patients. Yonsei Med J 2018 Mar;59(2):258-264 [FREE Full text] [doi: 10.3349/ymj.2018.59.2.258] [Medline: 29436194]
40. Blagus R, Lusa L. SMOTE for high-dimensional class-imbalanced data. BMC Bioinformatics 2013 Mar 22;14:106 [FREE Full text] [doi: 10.1186/1471-2105-14-106] [Medline: 23522326]
41. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res 2002 Jun 01;16:321-357. [doi: 10.1613/jair.953]
42. Lee T, Kim M, Kim S. Improvement of P300-based brain-computer interfaces for home appliances control by data balancing techniques. Sensors (Basel) 2020 Sep 29;20(19):5576 [FREE Full text] [doi: 10.3390/s20195576] [Medline: 33003367]

43. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition. Berlin, Germany: Springer; 2009.
44. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015 May 28;521(7553):436-444. [doi: 10.1038/nature14539] [Medline: 26017442]
45. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial intelligence in cardiology. J Am Coll Cardiol 2018 Jun 12;71(23):2668-2679 [FREE Full text] [doi: 10.1016/j.jacc.2018.03.521] [Medline: 29880128]
46. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589-1604 [FREE Full text] [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
47. Van Gelder I, Hagens V, Bosker H, Kingma JH, Kamp O, Kingma T, Rate Control versus Electrical Cardioversion for Persistent Atrial Fibrillation Study Group. A comparison of rate control and rhythm control in patients with recurrent persistent atrial fibrillation. N Engl J Med 2002 Dec 05;347(23):1834-1840. [doi: 10.1056/NEJMoa021375] [Medline: 12466507]
48. Roy D, Talajic M, Nattel S, Wyse DG, Dorian P, Lee KL, Atrial FibrillationCongestive Heart Failure Investigators. Rhythm control versus rate control for atrial fibrillation and heart failure. N Engl J Med 2008 Jun 19;358(25):2667-2677. [doi: 10.1056/NEJMoa0708789] [Medline: 18565859]
49. Carlsson J, Miketic S, Windeler J, Cuneo A, Haun S, Micus S, STAF Investigators. Randomized trial of rate-control versus rhythm-control in persistent atrial fibrillation: the Strategies of Treatment of Atrial Fibrillation (STAF) study. J Am Coll Cardiol 2003 May 21;41(10):1690-1696 [FREE Full text] [doi: 10.1016/s0735-1097(03)00332-2] [Medline: 12767648]
50. Wyse DG, Waldo AL, DiMarco JP, Domanski MJ, Rosenberg Y, Schron EB, Atrial Fibrillation Follow-up Investigation of Rhythm Management (AFFIRM) Investigators. A comparison of rate control and rhythm control in patients with atrial fibrillation. N Engl J Med 2002 Dec 05;347(23):1825-1833. [doi: 10.1056/NEJMoa021328] [Medline: 12466506]
51. Tsadok MA, Jackevicius CA, Essebag V, Eisenberg MJ, Rahme E, Humphries KH, et al. Rhythm versus rate control therapy and subsequent stroke or transient ischemic attack in patients with atrial fibrillation. Circulation 2012 Dec 04;126(23):2680-2687. [doi: 10.1161/CIRCULATIONAHA.112.092494] [Medline: 23124034]
52. Bunch T, Crandall B, Weiss J, May HT, Bair TL, Osborn JS, et al. Patients treated with catheter ablation for atrial fibrillation have long-term rates of death, stroke, and dementia similar to patients without atrial fibrillation. J Cardiovasc Electrophysiol 2011 Aug;22(8):839-845. [doi: 10.1111/j.1540-8167.2011.02035.x] [Medline: 21410581]
53. Noseworthy P, Gersh B, Kent D, Piccini JP, Packer DL, Shah ND, et al. Atrial fibrillation ablation in practice: assessing CABANA generalizability. Eur Heart J 2019 Apr 21;40(16):1257-1264 [FREE Full text] [doi: 10.1093/eurheartj/ehz085] [Medline: 30875424]
54. Kirchhof P, Bax J, Blomstrom-Lundquist C, Calkins H, Camm AJ, Cappato R, et al. Early and comprehensive management of atrial fibrillation: executive summary of the proceedings from the 2nd AFNET-EHRA consensus conference 'research perspectives in AF'. Eur Heart J 2009 Dec;30(24):2969-277c. [doi: 10.1093/eurheartj/ehp235] [Medline: 19535417]
55. Kirchhof P, Camm AJ, Goette A, Brandes A, Eckardt L, Elvan A, EAST-AFNET 4 Trial Investigators. Early rhythm-control therapy in patients with atrial fibrillation. N Engl J Med 2020 Oct 01;383(14):1305-1316. [doi: 10.1056/NEJMoa2019422] [Medline: 32865375]
56. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016 Jan 28;529(7587):484-489. [doi: 10.1038/nature16961] [Medline: 26819042]
57. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. Nature 2017 Oct 18;550(7676):354-359. [doi: 10.1038/nature24270] [Medline: 29052630]

Abbreviations

AF: atrial fibrillation
AUC: area under the receiver operator characteristic curve
EHR: electronic health record
ICD: International Classification of Disease
RF: random forest
SMOTE: synthetic minority oversampling technique
UC: University of Colorado

Edited by M Focsa; submitted 30.03.21; peer-reviewed by Z Ren, OS Liang, J Yang; comments to author 31.05.21; revised version received 15.07.21; accepted 11.08.21; published 06.12.21.

Please cite as:
Kim RS, Simon S, Powers B, Sandhu A, Sanchez J, Borne RT, Tumolo A, Zipse M, West JJ, Aleong R, Tzou W, Rosenberg MA
Machine Learning Methodologies for Prediction of Rhythm-Control Strategy in Patients Diagnosed With Atrial Fibrillation: Observational, Retrospective, Case-Control Study
JMIR Med Inform 2021;9(12):e29225
URL: https://medinform.jmir.org/2021/12/e29225
doi:10.2196/29225
PMID:34874889

Original Paper

Prediction Algorithms for Blood Pressure Based on Pulse Wave Velocity Using Health Checkup Data in Healthy Korean Men: Algorithm Development and Validation

Dohyun Park[1*], MA; Soo Jin Cho[2*], MD; Kyunga Kim[1,3*], PhD; Hyunki Woo[4], MA; Jee Eun Kim[2], MD; Jin-Young Lee[2], MD, PhD; Janghyun Koh[2], MD; JeanHyoung Lee[5], PhD; Jong Soo Choi[5], PhD; Dong Kyung Chang[1,6], MD, PhD; Yoon-Ho Choi[2], MD, PhD; Ji In Chung[2], MD; Won Chul Cha[1,5,7], MD; Ok Soon Jeong[5], MA; Se Yong Jekal[5], BSc; Mira Kang[1,2,5], MD, PhD

[1]Department of Digital Health, Samsung Advanced Institute of Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea

[2]Center for Health Promotion, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3]Statistics and Data Center, Research Institute for Future Medicine, Samsung Medical Center, Seoul, Republic of Korea

[4]Data Science Team, Evidnet Inc, Gyeonggi-do, Republic of Korea

[5]Digital Innovation Center, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[6]Division of Gastroenterology, Department of Internal Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[7]Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Mira Kang, MD, PhD
Department of Digital Health
Samsung Advanced Institute of Health Sciences and Technology
Sungkyunkwan University
81 Irwon-ro
Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 1099336838
Fax: 82 234101000
Email: mira90.kang@samsung.com

Abstract

Background: Pulse transit time and pulse wave velocity (PWV) are related to blood pressure (BP), and there were continuous attempts to use these to predict BP through wearable devices. However, previous studies were conducted on a small scale and could not confirm the relative importance of each variable in predicting BP.

Objective: This study aims to predict systolic blood pressure and diastolic blood pressure based on PWV and to evaluate the relative importance of each clinical variable used in BP prediction models.

Methods: This study was conducted on 1362 healthy men older than 18 years who visited the Samsung Medical Center. The systolic blood pressure and diastolic blood pressure were estimated using the multiple linear regression method. Models were divided into two groups based on age: younger than 60 years and 60 years or older; 200 seeds were repeated in consideration of partition bias. Mean of error, absolute error, and root mean square error were used as performance metrics.

Results: The model divided into two age groups (younger than 60 years and 60 years and older) performed better than the model without division. The performance difference between the model using only three variables (PWV, BMI, age) and the model using 17 variables was not significant. Our final model using PWV, BMI, and age met the criteria presented by the American Association for the Advancement of Medical Instrumentation. The prediction errors were within the range of about 9 to 12 mmHg that can occur with a gold standard mercury sphygmomanometer.

Conclusions: Dividing age based on the age of 60 years showed better BP prediction performance, and it could show good performance even if only PWV, BMI, and age variables were included. Our final model with the minimal number of variables (PWB, BMI, age) would be efficient and feasible for predicting BP.

(JMIR Med Inform 2021;9(12):e29212) doi:10.2196/29212

KEYWORDS

blood pressure; pulse transit time; pulse wave velocity; prediction model; algorithms; medical informatics; wearable devices

Introduction

High blood pressure (BP) is the leading cause of cardiovascular disease (CVD) such as coronary artery disease, stroke, heart failure, peripheral artery disease, and many kinds of microvascular disease. Furthermore, hypertension accounts for more CVD deaths than any other modifiable CVD risk factors. Most countries have published their own definition of hypertension and treatment guidelines. Those guidelines emphasize controlling BP in patients with hypertension because it can prevent CVD and reduce mortality according to a large amount of evidence [1-3]. For diagnosis and management of hypertension, accurate measurement of BP is crucial.

We can measure BP with many kinds of devices in an office setting and an out-of-office setting. However, BP varies with many factors such as cuff size and patient's position. Ambulatory BP monitoring with automated and programmable inflating cuff for 24 hours is considered as the reference standard BP since this method can rule out whitecoat hypertension or masked hypertension and measure nocturnal BP [4]. However, the aim of ambulatory BP is commonly diagnostic rather than real-time monitoring because the BP is measured in a fixed interval every 15 to 30 minutes over a 24-hour period. Several investigators tried to measure continuous BP using wearable devices with pulse transit time (PTT) and pulse wave velocity (PWV) to overcome disadvantages of ambulatory BP monitoring [5-10]. Although previous studies found a significant correlation between the PTT and the BP, they were conducted among a limited population of young and healthy male participants or among a small-sized population [11,12]. Therefore, they had limitations for generalization. To our knowledge, there was no investigation to evaluate the importance of each variable for prediction models as well.

The aim of this study is to develop BP prediction models with PWV in a large sample size of 1362 patients and to evaluate the relative importance of each clinical variable used in BP prediction models.

Methods

Study Population and Data Collection

This study was conducted on men older than 18 years who had a health medical examination at the Samsung Medical Center from January 2014 to December 2015 and conducted a test of the brachial-ankle PWV calculated by PTT. Among them, 1362 patients who were not taking antihypertensive medications or alpha-blockers for treating benign prostate hypertrophy were recruited for data analysis since these medications can affect PWV. Data was extracted from the Clinical Data Warehouse Darwin-C of Samsung Medical Center for this study. This study was approved by the Institutional Review Board (IRB) of the Samsung Medical Center (IRB number 2016-02-142). Each participant prepared a self-assessment questionnaire that included a past medical history, medication history, and smoking status. Smoking status was divided into three groups: nonsmokers, ex-smokers, and current smokers. Anthropometric measurements including body weight and height were performed with light clothing, and the BMI was calculated as weight (kg) divided by height (m^2) squared. Venous blood samples for high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides, glucose, hemoglobin A_{1c} (HbA_{1c}), creatinine, and C-reactive protein (CRP) were collected after 12-hour overnight fasting. Diabetes mellitus was defined as treated with diabetes medication, HbA_{1c}≥6.5%, or fasting glucose ≥126 mg/dL.

Pulse Wave Velocity and Blood Pressure

Brachial-ankle PWV was obtained using VP-1000 (Colin, Komaki, Japan) in the supine position with cuffs placed on both arms and ankles. They measure bilateral brachial and posterior tibial artery pressure waveform using an oscillometric method. PWV was calculated automatically with the distance from the heart to the ankle and the distance from the heart to the upper arm (L) divided by the pulse wave propagation time (PTT).

BP was obtained simultaneously with PWV measurement. Systolic BP (SBP) and diastolic BP (DBP) were defined as the average of pressures in both arms. Normotension was defined as SBP<140 mmHg and DBP<90 mmHg. Hypertension was defined as SBP≥140 mmHg or DBP≥90 mmHg.

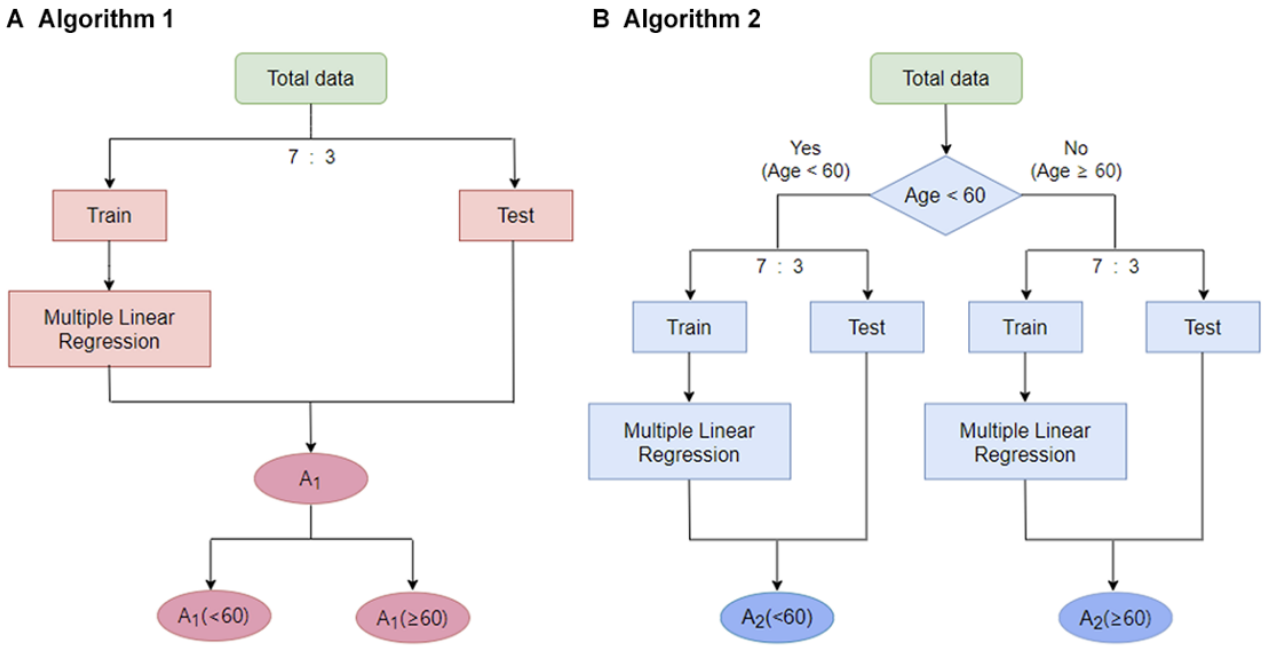
Statistical Analysis

Continuous variables were presented as means and SDs, and categorical variables were reported as percentages. Continuous variables were compared means between two groups using the Student *t* test, and categorical variables were compared frequencies through chi-square tests.

A total of 1362 participants were recruited in the model development cohort, and they were split up into the train and validation sets in a ratio of 7:3. The model development cohort repeated 200 different random seeds considering the effect of partition bias. The validation cohort was chosen randomly by selecting 100 patients each from the age group younger than 60 years and 60 years and older by stratifying age and BMI. Since BP and prostate medications can affect PWV, these patients were excluded. After Spearman correlation was conducted for 33 variables, including PWV, age, questionnaires, physical

information, and chemistry electrolyte tests, 17 variables with absolute values of correlation numbers of 0.5 or less were selected to exclude multicollinearity. The BP prediction model used multilinear regression analysis, and this study compared the performance of the model based on the total age (algorithm 1) and the model made by dividing it into two age groups based on the age of 60 years (algorithm 2; Figure 1).

Figure 1. Algorithms based on subgroups by age.



Model 1 was done by only using PWV. Model 2 was made by adding BMI and age to model 1. Model 3 is a nested model that includes heart rate (HR), smoke status, white blood cell count (WBC), hemoglobin, uric acid, sodium, potassium, LDL, HDL, triglyceride, testosterone, creatinine, CRP, and diabetes into model 2. Two sample *t* tests were used for the comparison between the two groups, and differences from zero were compared through one sample *t* test. Analysis of variance was used to compare the performance between the three models. For the post hoc test, the most conservative Bonferroni test was used. We used Johnson relative weights to quantify the relative importance of correlated predictor variables in multiple linear regression analysis [13]. For the evaluation of the performance of the BP prediction model, error was used to indicate the difference between the predicted value and the actual value, and root mean squared error (RMSE) to indicate the predicted error of the continuous variable. The RMSE is defined as the square root of the mean of the difference between the predicted and the real value. The final model was evaluated for the performance of the model compared to the BP medical device grading criteria suggested by the British Hypertension Society (BHS) and the American Association for the Advancement of Medical Instrumentation (AAMI) [14,15]. All analyses determined statistical significance based on the significance level of .05. For statistical analysis, R 4.02 version (R Foundation for Statistical Computing) was used.

Results

Baseline Characteristics

Based on the data of 1362 adult males older than 18 years in this study, the baseline clinical characteristics of study participants are shown in Table 1. Participants were aged between 18 and 90 years, with an average of 62.1 (SD 7.7) years. Of the 1362 people, 303 were younger than 60 years, while 1059 were older than 60 years. The normal BP was 1117, and the high pressure was 245. People with hypertension had higher PWV, BMI, HR, WBC, HDL, triglyceride, uric acid, and testosterone than normal people.

Table 1. Baseline clinical characteristics of study participants.

Characteristic[a]	Age groups			BP[b] groups			All (N=1362)
	Age <60 years (n=303)	Age≥60 years (n=1059)	*P* value	Normal BP (n=1117)	Hypertension (n=245)	*P* value	
Age (years)	50.9 (4.9)	65.3 (4.9)	<.001	62.2 (7.5)	61.7 (8.6)	.39	62.1 (7.7)
SBP[c] (mmHg)	124.3 (12.8)	125.1 (13.2)	.36	120.7 (9.4)	143.9 (10.5)	<.001	124.9 (13.1)
DBP[d] (mmHg)	81.5 (9.3)	79.1 (7.9)	<.001	77.2 (6.6)	90.6 (6.0)	<.001	79.6 (8.3)
PWV[e] average (cm/s)	1378.2 (157.6)	1544.4 (257.4)	<.001	1466.7 (217.4)	1692.8 (293.9)	< .001	1507.4 (248.6)
BMI (kg/m^2)	24.6 (2.6)	24.0 (2.5)	<.001	23.9 (2.4)	24.9 (2.8)	<.001	24.1 (2.5)
Heart rate (BPM)	63.3 (9.5)	63.1 (9.9)	.73	62.6 (9.6)	65.3 (10.5)	<.001	63.1 (9.8)
White blood cell count (10^3/μL)	5.7 (1.6)	5.7 (1.6)	.82	5.6 (1.5)	6.0 (1.6)	<.001	5.7 (1.6)
Hemoglobin (g/dL)	15.4 (1.0)	15.1 (1.1)	<.001	15.1 (1.1)	15.3 (1.2)	.12	15.2 (1.1)
Uric acid (mg/dL)	5.9 (1.3)	5.6 (1.2)	<.001	5.6 (1.2)	5.9 (1.3)	.001	5.7 (1.2)
Sodium (mEq/L)	142.0 (1.7)	142.1 (1.8)	.30	142.1 (1.8)	142.1 (1.9)	.62	142.1 (1.8)
Potassium (mEq/L)	4.4 (0.3)	4.4 (0.3)	.58	4.4 (0.3)	4.4 (0.4)	.32	4.4 (0.3)
Low-density lipoprotein (mg/dL)	126.7 (28.5)	117.2 (31.0)	<.001	119.1 (30.9)	120.4 (30.0)	.56	119.3 (30.7)
High-density lipoprotein (mg/dL)	55.4 (14.5)	55.1 (14.3)	.74	55.6 (14.6)	53.0 (12.9)	.007	55.1 (14.4)
Triglyceride (mg/dL)	124.6 (74.0)	113.6 (64.8)	.02	113.0 (62.1)	129.8 (85.1)	.004	116.0 (67.1)
Testosterone (ng/mL)	5.3 (1.5)	5.2 (1.7)	.77	5.3 (1.6)	4.9 (1.7)	<.001	5.2 (1.6)
Smoking status, n (%)			<.001			.55	
Never smoker	73 (24.1)	280 (26.4)		290 (26.0)	63 (25.7)		353 (25.9)
Ex-smoker	143 (47.2)	589 (55.6)		594 (53.2)	138 (56.3)		732 (53.7)
Current smoker	87 (28.7)	190 (17.9)		233 (29.0)	44 (18.0)		277 (20.3)
Creatinine (mg/dL)	1.0 (0.1)	1.0 (0.3)	.78	1.0 (0.1)	1.0 (0.5)	.14	1.0 (0.2)
C-reactive protein (mg/dL)	0.1 (0.3)	0.1 (0.3)	.46	0.1 (0.3)	0.1 (0.3)	.25	0.1 (0.3)
Diabetes, n (%)			<.001			.06	
No	274 (90.4)	853 (80.5)		192 (78.4)	192 (78.4)		1127 (82.7)
Yes	29 (9.6)	206 (19.5)		53 (21.6)	53 (21.6)		235 (17.3)

[a]Values are reported as mean (SD).

[b]BP: blood pressure.

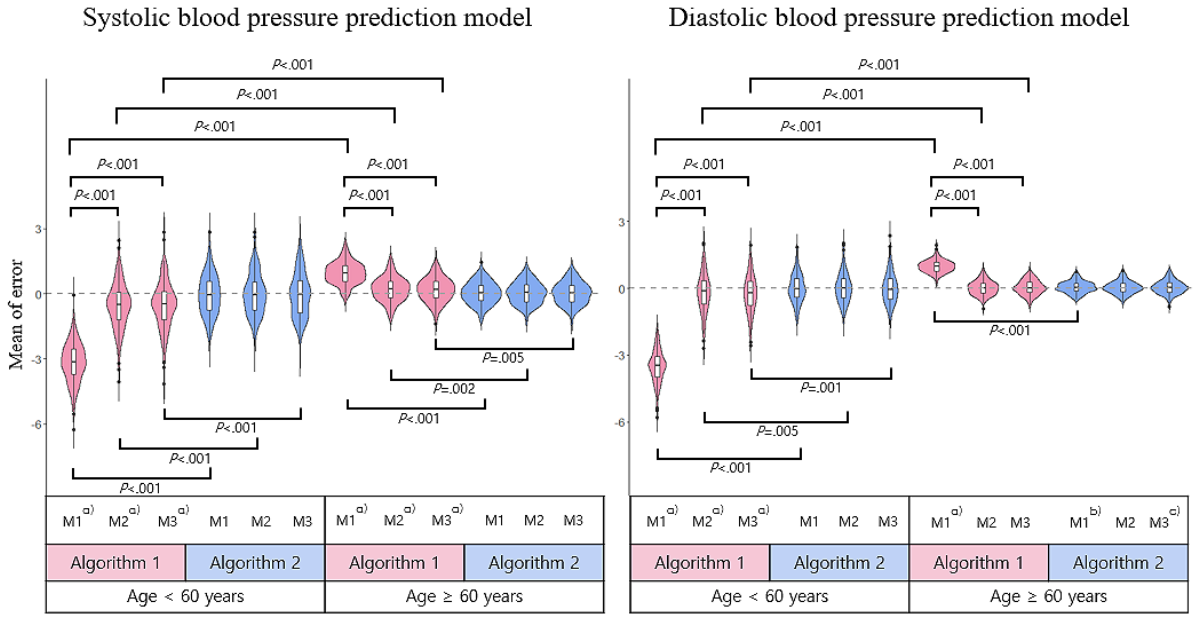
[c]SBP: systolic blood pressure.

[d]DBP: diastolic blood pressure.

[e]PWV: pulse wave velocity.

Figure 2 shows the distribution for the mean of error obtained from repeated analysis of 200 random seeds. When estimating SBP with algorithm 1, underestimation occurred in the younger than 60 years age group, and overestimation occurred in the 60 years and older group (*P*<.001, one-sample *t* test). When estimating DBP with algorithm 1, underestimation occurred in the younger than 60 years group, and overestimation was only performed on model 1 for those 60 years and older. In algorithm 1, model 1 had the worst performance. The average of the mean of error was significantly smaller when algorithm 2 was applied in SBP forecasts compared to algorithm 1. In the case of DBP prediction, the average of the mean of error was significantly lower when algorithm 2 was applied in comparison to algorithm 1 in the younger than 60 years age group.

Figure 2. Blood pressure prediction model: mean of error of blood pressure based on 200 repetitive partitions.



M1=Model 1; M2=Model 2; M3=Model 3
One sample *t* test (mean comparison with zero): a) *P*<.001, b) *P*=.01, c) *P*=.04

From the view of RMSE, the performance of model 2 with the condition of age <60 years was better in algorithm 2 than in algorithm 1. The performance of model 2 with the condition of age ≥60 years was not significantly different between algorithms 1 and 2 (Figure 3). When comparing between models in algorithm 2, model 1 was the worst, and there was no significant difference in performance between model 2 and model 3 (Table 2).

Figure 3. The distribution of RMSE in model 2 obtained from repeating the analysis by 200 random seeds. RMSE: root mean square error.

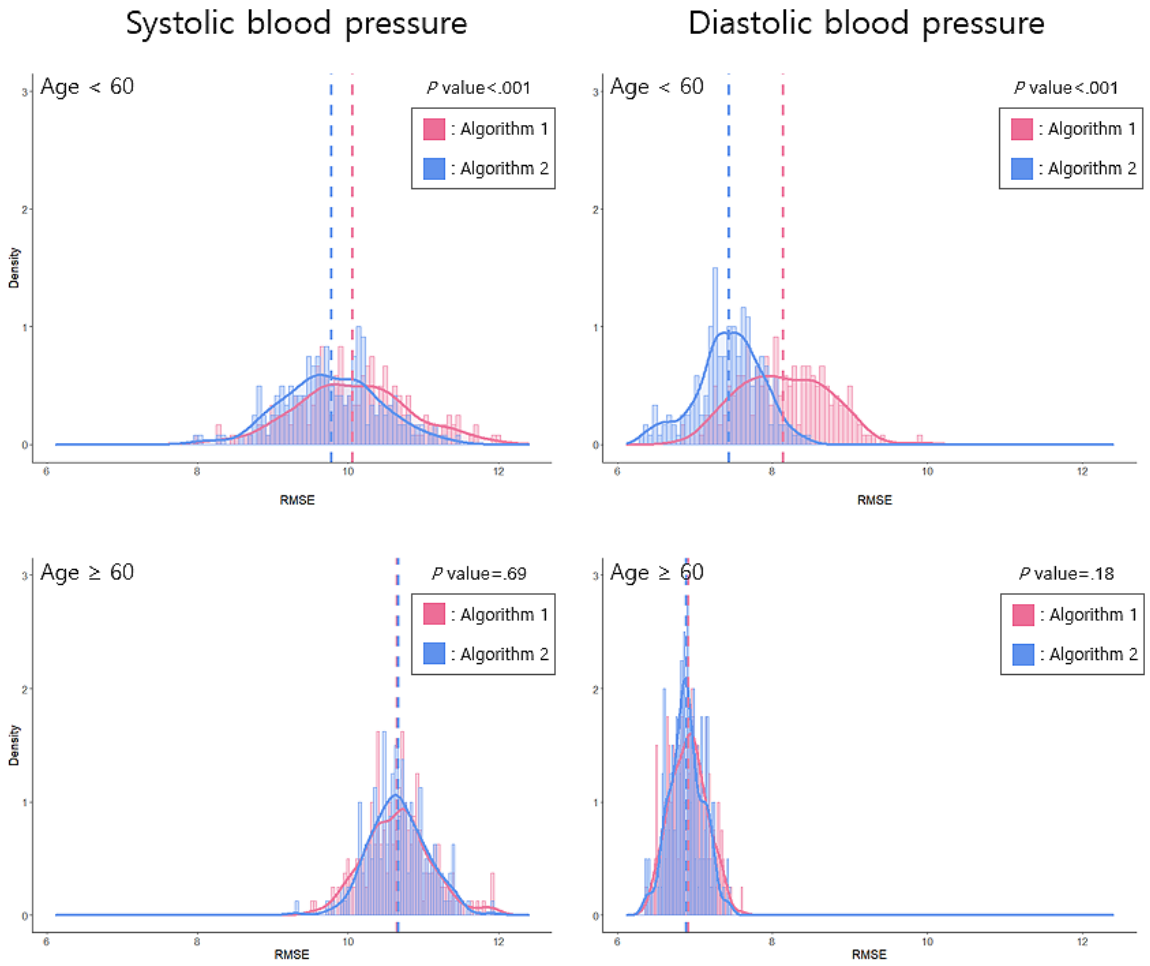


Table 2. RMSE[a] of the models in algorithm 2.

Models	Systolic blood pressure		Diastolic blood pressure	
	RMSE of A_2 (<60 years; SD)	RMSE of A_2 (≥60 years; SD)	RMSE of A_2 (<60 years; SD)	RMSE of A_2 (≥60 years; SD)
Model 1	10.31 (0.67)	10.88 (0.4)	7.74 (0.45)	7.21 (0.2)
Model 2	9.78 (0.63)	10.67 (0.38)	7.43 (0.43)	6.88 (0.21)
Model 3	9.99 (0.66)	10.61 (0.39)	7. 33 (0.44)	6.76 (0.22)

[a]RMSE: root mean square error.

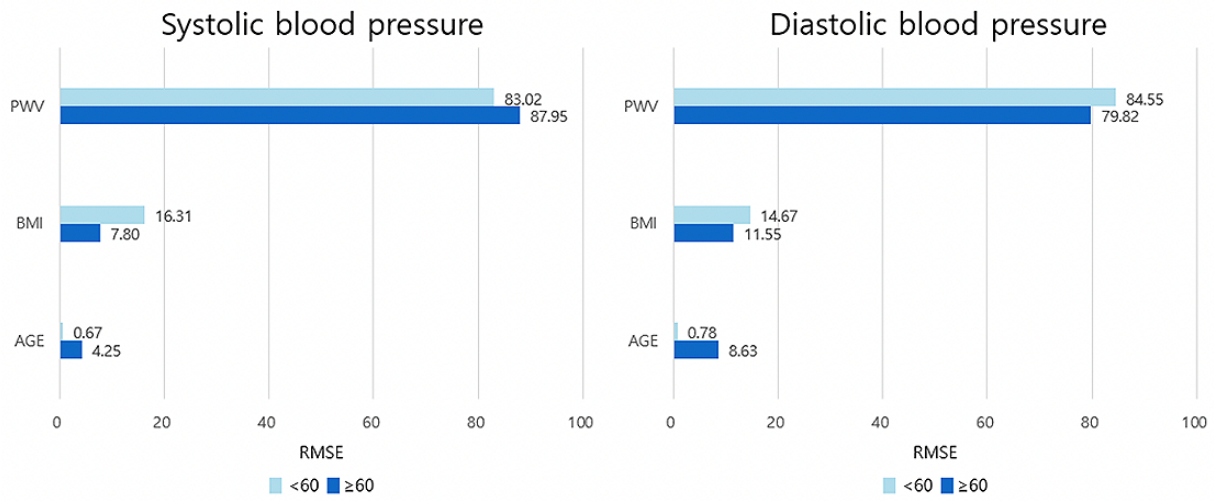
After considering all the aforementioned, we selected model 2 based on algorithm 2 as the best prediction model. Table 3 shows the final prediction equation of the multiple linear regression model. SBP and DBP are in direct proportion to PWV and BMI. The influence of PWV on SBP and DBP was more apparent in those aged <60 years than in those aged ≥60 years, so was BMI (Table 3). PWV contributed the most to BP prediction, followed by BMI and age (Figure 4). Model 3, which used 17 variables, also had the greatest influence of PWV (Multimedia Appendix 1).

Table 3. Final prediction equation from model 2 built in algorithm 2.

Blood pressure and variables	Age <60 years		Age ≥60 years	
	Nonstandardized regression coefficient	*P* value	Nonstandardized regression coefficient	*P* value
Systolic blood pressure				
Constant	29.9756	.002	68.3969	<.001
PWV[a] average (cm/s)	0.0487	<.001	0.0304	<.001
Age (years)	–0.0882	.46	–0.1669	.02
BMI (kg/m^2)	1.2876	<.001	0.8606	<.001
Diastolic blood pressure				
Constant	11.4629	.12	69.9290	<.001
PWV average (cm/s)	0.0322	<.001	0.0149	<.001
Age (years)	0.0653	.47	–0.3990	<.001
BMI (kg/m^2)	0.9057	<.001	0.5076	<.001

[a]PWV: pulse wave velocity.

Figure 4. Relative explanatory power (R^2) between the variables of the final model in the model development cohort. PWV: pulse wave velocity; RMSE: root mean square error.



Assessment for the Performance of BP Prediction

To evaluate the performance of the final prediction model, criteria provided by the AAMI and the BHS were applied. All of AAMI's criteria were satisfied, and BHS's criteria were only met by the 60 years or older DBP with class A. Although our prediction model did not meet the BHS criteria, it is still within acceptable range for clinical use according to AAMI's protocol (Table 4).

Table 4. AAMI[a] and BHS[b] grading of models with the data divided into three pressure categories.

Category and grade	AAMI[c] mean difference between standard and test device (mmHg), absolute mean difference (SD)	Grade	BHS[d] absolute difference between standard and test device (mmHg)		
			≤5	≤10	≤15
Grading criteria					
Passed	≤5 (≤8)	A	60%	85%	95%
Passed	≤5 (>8)	B	50%	75%	90%
Passed	>5 (≤8)	C	40%	65%	85%
Failed	>5 (>8)	D	—[e]	—	—
Age <60 years					
SBP[f] (passed)	2.25 (7.69)	C	43%	83%	93%
DBP[g] (passed)	3.05 (6.07)	C	49%	85%	99%
Age ≥60 years					
SBP (passed)	1.33 (9.73)	C	41%	66%	87%
DBP (passed)	0.09 (6.63)	A	60%	89%	98%

[a]AAMI: Association for the Advancement of Medical Instrumentation.

[b]BHS: British Hypertension Society.

[c]To meet AAMI criteria, the mean difference between the device and the mercury standard must be ≤5 mmHg or the SD must be ≤8 mmHg.

[d]To meet BHS criteria, devices must achieve a grade of at least B for both systolic and diastolic measurements. Grade A denotes greatest agreement with mercury standard and D denotes least agreement.

[e]Worse than a C.

[f]SBP: systolic blood pressure.

[g]DBP: diastolic blood pressure.

Discussion

Principal Findings

About 30% of the world's deaths are caused by CVD [16]. Among the risk factors for CVD, high BP is one of the most common causes of premature cardiovascular death, but it is modifiable [17]. Every 10 mmHg reduction of SBP can reduce the risk of major CVD events: 17% reduction in coronary heart disease, 27% reduction in stroke, 28% reduction in heart failure, and 13% reduction in all-cause mortality [18]. All global guidelines recommend strict control of BP, and the accurate measurement of BP is the first step in BP management.

Most people measure BP in the office, but the office BP is relatively inaccurate compared to other measurement methods due to many factors such as cuff size, patient's position, and emotional state. Therefore, recent guidelines recommend other methods of BP measurement such as ambulatory or home BP monitoring [19,20]. However, ambulatory BP monitoring is not easy to obtain since not all clinics have special devices. In addition, it is not comfortable for patients to cover their upper arms for a 24-hour duration with programmed inflating cuff in daily life. Home BP monitoring can obtain more accurate values than office BP because it is measured in stable states in most cases. However, there is still a limitation in getting continuous BP.

Recently, continuous BP monitoring with PTT and PWV was developed to compensate for the weaknesses of conventional BP measurement methods. Many attempts have been made using wearable devices attached to chest, ear, or wrist for continuous monitoring. However, previous studies were small in a sample size of less than 500 patients, and there was no study to evaluate the relative importance of clinical variables in predicting BP. We made BP predicting models using PWV and clinical data based on a large-scale population of over a thousand and evaluated the relative importance of the clinical variables.

After creating various types of BP predicting models, we concluded that the performance of the models was better in age-based stratification since the cardiovascular system changes as the age increases. The prevalence of hypertension is 30% to 45% in adults, and hypertension becomes progressively more common with age. Over 60% of people aged older than 60 years are diagnosed with hypertension [1]. Moreover, with or without hypertension, SBP and DBP tend to change differently with aging. DBP tends to increase until the age of 60 years and decrease after this age, but SBP increases continuously even after the age of 60 years [21]. This phenomenon is attributed to increasing stiffness of aortic wall caused by changing inert elastic fibers. Increased stiffness of aortic wall results in increase in PWV. Increase in PWV causes early reflection of pulse from peripheral arterioles and augments pressure in late systole rather than early diastole. This explains the constant increase in SBP and decrease in DBP in those aged around 60 years [22]. One of the previous SBP prediction models showed better

performance when it was divided into age groups of younger than 60 years and older than 60 years [23]. We created models for both SBP and DBP separately in consideration of natural vascular aging. Our prediction model for both SBP and DBP had better performance when using algorithm 2, which was stratified by the age of 60 years.

Although the exact etiology of primary hypertension remains unclear, a number of risk factors are strongly and independently associated with its development, including not only age but also race, family history, obesity, diet, and physical activity [24,25]. In addition, many studies showed modifiable risk factors for CVD such as smoking, diabetes mellitus, dyslipidemia, and obesity, which are common in adults with hypertension because these risk factors and hypertension share the mechanism of pathophysiology. They found these risk factors affect BP through overactivation of the renin angiotensin aldosterone system and sympathetic nervous system, inhibition of the cardiac natriuretic peptide system, and endothelial dysfunction. Therefore, modification of cardiovascular risk factors may affect BP [3]. We made model 3 using 17 variables including clinical information. At the beginning, we expected model 3 would be more accurate than model 2, as model 2 was nested from model 3. However, there was no significant difference in performance between model 2 and model 3. For that reason, we adopted model 2 for convenience because BMI and age are easy to obtain in daily life.

Our final model, model 2 in algorithm 2, satisfied the criteria of the AAMI by the mean of error, although it did not meet the criteria of the BHS in absolute pressure difference. The prediction errors were within the range of about 9 to 12 mmHg that can occur with a gold standard mercury sphygmomanometer. According to a previous validation survey by O'Brian et al [26], only a few BP measuring devices met the standards in both criteria. This study validated 21 commercially available devices for the self-measurement of BP. Some BP measuring devices were in grade D in the BHS standard, and only five devices satisfied both standards [26]. Therefore, our prediction model can be useful in practice.

In conclusion, stratification of age is important in developing a BP prediction model with better accuracy. In addition, BP is influenced predominantly by PWV, BMI, and age out of other clinical factors. Our final model with minimal number of variables would be efficient and feasible for predicting BP.

Limitation

This analysis was conducted among healthy male participants. The study population included patients that were hypotensive and hypertensive but excluded those taking antihypertensive drugs. Further studies should be warranted on a diverse population, including patients on antihypertensive medications and female participants, and on the performance of PWV in wider range of BPs.

The Health Promotion Center at Samsung Medical Center does not request detailed medication information except for hypertensive medication on the personal questionnaire for health checkup. Receiving additional information on medication is impossible, as this is a retrospective study. Accordingly, there is some limitation in analyzing the effects of different types of medication such as alpha-blockers or calcium channel blockers. Further studies are needed including drug information.

Our prediction model was internally validated; however, this model should be validated externally.

Acknowledgments

This study was supported by a National IT Industry Promotion Agency grant funded by the Ministry of Science and ICT, and Ministry of Health and Welfare (project S1906-21-1001; Development Project of the Precision Medicine Hospital Information System). This study was also supported by the Technology Innovation Program (program 20005021: Establishment of Standardization and Anonymization Guidelines Based on a Common Data Model; program 20011642: common data model–based algorithm for treatment protocol service system development and spread), which was funded by the Ministry of Trade, Industry and Energy in Korea.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Relative explanatory power (R^2) between the 17 variables in the model development cohort.
[DOCX File , 119 KB - medinform_v9i12e29212_app1.docx]

References

1. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M, ESC Scientific Document Group. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J 2018 Sep 01;39(33):3021-3104. [doi: 10.1093/eurheartj/ehy339] [Medline: 30165516]
2. Lee H, Shin J, Kim G, Park S, Ihm S, Kim HC, et al. 2018 Korean Society of Hypertension Guidelines for the management of hypertension: part II-diagnosis and treatment of hypertension. Clin Hypertens 2019;25:20 [FREE Full text] [doi: 10.1186/s40885-019-0124-x] [Medline: 31388453]

3. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Dennison Himmelfarb C, et al. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. J Am Coll Cardiol 2018 May 15;71(19):e127-e248 [FREE Full text] [doi: 10.1016/j.jacc.2017.11.006] [Medline: 29146535]
4. Muntner P, Shimbo D, Carey RM, Charleston JB, Gaillard T, Misra S, et al. Measurement of blood pressure in humans: a scientific statement from the American Heart Association. Hypertension 2019 May;73(5):e35-e66 [FREE Full text] [doi: 10.1161/HYP.0000000000000087] [Medline: 30827125]
5. Park J, Yang S, Sohn J, Lee J, Lee S, Ku Y, et al. Cuffless and continuous blood pressure monitoring using a single chest-worn device. IEEE Access 2019;7:135231-135246. [doi: 10.1109/access.2019.2942184]
6. Solá J, Proença M, Chételat O. Wearable PWV technologies to measure blood pressure: eliminating brachial cuffs. Annu Int Conf IEEE Eng Med Biol Soc 2013;2013:4098-4101. [doi: 10.1109/EMBC.2013.6610446] [Medline: 24110633]
7. Holz C, Wang EJ. Glabella: continuously sensing blood pressure behavior using an unobtrusive wearable device. Proc ACM Interactive Mobile Wearable and Ubiquitous Technologies 2017 Sep 11;1(3):1-23. [doi: 10.1145/3132024]
8. Carek AM, Conant J, Joshi A, Kang H, Inan OT. SeismoWatch: wearable cuffless blood pressure monitoring using pulse transit time. Proc ACM Interact Mob Wearable Ubiquitous Technol 2017 Sep;1(3):40 [FREE Full text] [doi: 10.1145/3130905] [Medline: 30556049]
9. Heartisans. URL: https://www.heartisans.com/ [accessed 2020-08-10]
10. Carek A, Holz C. Naptics: convenient and continuous blood pressure monitoring during sleep. Proc ACM Interactive Mobile Wearable Ubiquitous Technologies 2018 Sep 18;2(3):1-22. [doi: 10.1145/3264906]
11. Gesche H, Grosskurth D, Küchler G, Patzak A. Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method. Eur J Appl Physiol 2012 Jan;112(1):309-315. [doi: 10.1007/s00421-011-1983-3] [Medline: 21556814]
12. Wong MY, Poon CC, Zhang Y. An evaluation of the cuffless blood pressure estimation based on pulse transit time technique: a half year study on normotensive subjects. Cardiovasc Eng 2009 Mar;9(1):32-38. [doi: 10.1007/s10558-009-9070-7] [Medline: 19381806]
13. Thomas D, Zumbo B, Kwan E, Schweitzer L. On Johnson's (2000) relative weights method for assessing variable importance: a reanalysis. Multivariate Behav Res 2014;49(4):329-338. [doi: 10.1080/00273171.2014.905766] [Medline: 26765801]
14. American National Standards Institute. American National Standard for Electronic Or Automated Sphygmomanometers. Arlington, VA: AAMI; 1987:1-25.
15. O'Brien E, Petrie J, Littler W, de Swiet M, Padfield PL, O'Malley K, et al. The British Hypertension Society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems. J Hypertens 1990 Jul;8(7):607-619. [doi: 10.1097/00004872-199007000-00004] [Medline: 2168451]
16. Chopra H, Ram CVS. Recent guidelines for hypertension. Circ Res 2019 Mar 29;124(7):984-986. [doi: 10.1161/circresaha.119.314789]
17. Tackling G, Borhade M. Hypertensive Heart Disease. Treasure Island, FL: StatPearls; Jan 2021.
18. Ettehad D, Emdin CA, Kiran A, Anderson SG, Callender T, Emberson J, et al. Blood pressure lowering for prevention of cardiovascular disease and death: a systematic review and meta-analysis. Lancet 2016 Mar 05;387(10022):957-967. [doi: 10.1016/S0140-6736(15)01225-8] [Medline: 26724178]
19. Daskalopoulou SS, Rabi DM, Zarnke KB, Dasgupta K, Nerenberg K, Cloutier L, et al. The 2015 Canadian Hypertension Education Program recommendations for blood pressure measurement, diagnosis, assessment of risk, prevention, and treatment of hypertension. Can J Cardiol 2015 May;31(5):549-568. [doi: 10.1016/j.cjca.2015.02.016] [Medline: 25936483]
20. Parati G, Stergiou G, O'Brien E, Asmar R, Beilin L, Bilo G, European Society of Hypertension Working Group on Blood Pressure Monitoring and Cardiovascular Variability. European Society of Hypertension practice guidelines for ambulatory blood pressure monitoring. J Hypertens 2014 Jul;32(7):1359-1366. [doi: 10.1097/HJH.0000000000000221] [Medline: 24886823]
21. Franklin SS, Gustin W, Wong ND, Larson MG, Weber MA, Kannel WB, et al. Hemodynamic patterns of age-related changes in blood pressure. The Framingham Heart Study. Circulation 1997 Jul 01;96(1):308-315. [doi: 10.1161/01.cir.96.1.308] [Medline: 9236450]
22. O'Rourke MF, Nichols WW. Aortic diameter, aortic stiffness, and wave reflection increase with age and isolated systolic hypertension. Hypertension 2005 Apr;45(4):652-658. [doi: 10.1161/01.HYP.0000153793.84859.b8] [Medline: 15699456]
23. Suzuki S, Oguri K. Cuffless and non-invasive systolic blood pressure estimation for aged class by using a photoplethysmograph. 2008 Presented at: 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society; August 20-25, 2008; Vancouver, BC p. 1327-1330. [doi: 10.1109/iembs.2008.4649409]
24. Forman JP, Stampfer MJ, Curhan GC. Diet and lifestyle risk factors associated with incident hypertension in women. JAMA 2009 Jul 22;302(4):401-411 [FREE Full text] [doi: 10.1001/jama.2009.1060] [Medline: 19622819]
25. Wang N, Young JH, Meoni LA, Ford DE, Erlinger PT, Klag JM. Blood pressure change and risk of hypertension associated with parental hypertension: the Johns Hopkins Precursors Study. Arch Intern Med 2008 Mar 24;168(6):643-648. [doi: 10.1001/archinte.168.6.643] [Medline: 18362257]

26. O'Brien E, Waeber B, Parati G, Staessen J, Myers M. Blood pressure measuring devices: recommendations of the European Society of Hypertension. BMJ 2001 Mar 03;322(7285):531-536 [FREE Full text] [doi: 10.1136/bmj.322.7285.531] [Medline: 11230071]

Abbreviations

AAMI: American Association for the Advancement of Medical Instrumentation
BHS: British Hypertension Society
BP: blood pressure
CRP: C-reactive protein
CVD: cardiovascular disease
DBP: diastolic blood pressure
HbA_{1c}: hemoglobin A_{1c}
HDL: high-density lipoprotein
HR: heart rate
IRB: Institutional Review Board
LDL: low-density lipoprotein
PTT: pulse transit time
PWV: pulse wave velocity
RMSE: root mean squared error
SBP: systolic blood pressure
WBC: white blood cell count

Edited by C Lovis; submitted 30.03.21; peer-reviewed by K Ho; comments to author 25.07.21; revised version received 06.08.21; accepted 24.09.21; published 08.12.21.

Please cite as:
Park D, Cho SJ, Kim K, Woo H, Kim JE, Lee JY, Koh J, Lee J, Choi JS, Chang DK, Choi YH, Chung JI, Cha WC, Jeong OS, Jekal SY, Kang M
Prediction Algorithms for Blood Pressure Based on Pulse Wave Velocity Using Health Checkup Data in Healthy Korean Men: Algorithm Development and Validation
JMIR Med Inform 2021;9(12):e29212
URL: https://medinform.jmir.org/2021/12/e29212
doi:10.2196/29212
PMID:34889753

Original Paper

On Missingness Features in Machine Learning Models for Critical Care: Observational Study

Janmajay Singh[1], BTECH; Masahiro Sato[1], PhD; Tomoko Ohkuma[1], PhD

Fuji Xerox Co, Ltd, Yokohama, Japan

Corresponding Author:
Janmajay Singh, BTECH
Fuji Xerox Co, Ltd
6 Chome-1-1 Minatomirai
Nishi Ward
Yokohama, 220-0012
Japan
Phone: 81 7041120526
Email: janmajaysingh14@gmail.com

Abstract

Background: Missing data in electronic health records is inevitable and considered to be nonrandom. Several studies have found that features indicating missing patterns (missingness) encode useful information about a patient's health and advocate for their inclusion in clinical prediction models. But their effectiveness has not been comprehensively evaluated.

Objective: The goal of the research is to study the effect of including informative missingness features in machine learning models for various clinically relevant outcomes and explore robustness of these features across patient subgroups and task settings.

Methods: A total of 48,336 electronic health records from the 2012 and 2019 PhysioNet Challenges were used, and mortality, length of stay, and sepsis outcomes were chosen. The latter dataset was multicenter, allowing external validation. Gated recurrent units were used to learn sequential patterns in the data and classify or predict labels of interest. Models were evaluated on various criteria and across population subgroups evaluating discriminative ability and calibration.

Results: Generally improved model performance in retrospective tasks was observed on including missingness features. Extent of improvement depended on the outcome of interest (area under the curve of the receiver operating characteristic [AUROC] improved from 1.2% to 7.7%) and even patient subgroup. However, missingness features did not display utility in a simulated prospective setting, being outperformed (0.9% difference in AUROC) by the model relying only on pathological features. This was despite leading to earlier detection of disease (true positives), since including these features led to a concomitant rise in false positive detections.

Conclusions: This study comprehensively evaluated effectiveness of missingness features on machine learning models. A detailed understanding of how these features affect model performance may lead to their informed use in clinical settings especially for administrative tasks like length of stay prediction where they present the greatest benefit. While missingness features, representative of health care processes, vary greatly due to intra- and interhospital factors, they may still be used in prediction models for clinically relevant outcomes. However, their use in prospective models producing frequent predictions needs to be explored further.

(JMIR Med Inform 2021;9(12):e25022) doi:10.2196/25022

KEYWORDS

electronic health records; informative missingness; machine learning; missing data; hospital mortality; sepsis

Introduction

Background

The increasing availability of electronic health record (EHR) data collected from hospitals, especially from their intensive care units (ICU), has encouraged the development of various models for disease diagnosis [1-4]. Machine learning and specifically deep learning models, given their ability to adequately learn nonlinear representations and temporal patterns from large amounts of data, have been widely applied to capture complex physiological processes, and several works have demonstrated their usefulness [5]. Most works use retrospective observational data to train supervised models for a variety of

clinically important outcomes like mortality or sepsis. Some more recent works have also developed models more suited to actual clinical needs by evaluating models prospectively and using early warning scores as baselines [6]. Models used to learn human physiological processes from EHRs tackle intrinsic problems in health care data, particularly that of irregular sampling and large amount of missing information [7].

Several methods have been developed to handle the inevitably large amount of missing data in EHRs. Simpler methods like incomplete record deletion (also called complete case analysis) propose to simply delete those records where any value is missing. Various imputation techniques ranging from simple mean imputation to sophisticated methods like multiple imputation with chained equations are also commonly used [8]. More recently, deep learning models have been proposed to learn the underlying process generating the data as a method for better inferring missing values [9]. A consensus regarding a best universal model to handle missing data does not exist in literature, and it is generally understood to depend heavily on the task and the nature of the data itself. However, a returning consideration in all studies on missing data is the nature of missingness. In Rubin [10], missing data were classified into 3 categories: missing completely at random, missing at random, and missing not at random. The nature of missingness in EHRs has been generally understood to belong to the last category, missing not at random [11]. This means that missing values cannot be inferred using observed values, subjecting all methods to problems of bias.

Considering the inevitability of bias, methods seek to minimize it by considering imputed value uncertainty or developing more sophisticated processes to learn underlying distributions [8,12]. A returning simple yet effective motif in deep learning models for EHRs is to use informative missingness (IM) features. First introduced in Lin and Haug [11], the method has repeatedly been shown to improve performance of health care models for a variety of outcomes [13-16]. A particularly efficient use was demonstrated in Lipton et al [13], where simply augmenting zero-imputed data with corresponding binary missingness indicators greatly improved over the baseline model. The basic assumption underlying the use of IM features is that the inclusion of health care process variables like laboratory tests conducted or drugs prescribed provides important information about the state and evolution of a patient's health. These variables are usually inputted to the model as binary indicators of observation/missingness, but some studies have also propounded modifying or augmenting this representation to include additional information such as time since last observation [17,18]. We use the term health care process variables interchangeably with IM features.

This use of health care process variables as feasible features to model patient health is supported by studies spanning several decades and countries, indicating that test ordering behavior and drug prescriptions are associated with the underlying pathology. For example, Kristiansen et al [19] established that the medical condition at hand was the strongest determinant of test ordering behavior, and Weiskopf et al [20] and Rusanov et al [21] found a statistically significant relationship between data completeness and patient health status, finding that those susceptible to adverse outcomes have more information collected. A recent study also highlighted that EHR data are observational and display a patient's interactions with the health care system and thus any information from there can only serve as a proxy measure of the patient's true state [22]. They further found that the presence of laboratory test orders, regardless of other information like numerical test values, had a significant association with odds of 3-year survival. This suggests that laboratory test orders encode information separately from laboratory test results, as corroborated by Pivovarov et al [23].

Despite improvements in model performance on including IM features, their use is considered to have limited applicability. Missing information may occur due to several factors, not all which pertain to patient pathology or a physician's mental model of the diagnosis process. Within a hospital, some tests may be conducted following general guidelines or as standard practice for all patients regardless of underlying condition [23]. Physicians also vary by years of experience and attitudes in coping with uncertainty, which has been shown to affect test ordering behavior [24]. In addition, variations between hospitals as test ordering may depend on resource constraints and variations due to geographic separation as ICU case-mix changes are further exacerbated when making international comparisons [25,26]. And while machine learning models rely on improved performance on chosen metrics as a justification for continued use of IM features, evaluation has mostly been on single-center data under retrospective task settings. Even where multicenter data are used, hospitals are often not geographically distinct, preventing the assessment of model generalization to different demographic mixes and practices. Also, only recently have some works evaluated their models prospectively, better reflecting real-world clinical utility, but evaluation metrics differ across studies, some choosing to use the concordance index (also called the area under receiver operating curve [AUROC]) while others prefer the area under precision recall curve [27,28].

The ways in which use of IM features is supported and challenged creates an apparent disjunction and casts doubts on their true usefulness. This was perhaps exemplified in the PhysioNet 2019 Challenge [29] for early prediction of sepsis, which saw many submissions using some modification of IM features [16-18,30,31]. The challenge was designed to evaluate models on prospective prediction performance and used datasets from 3 geographically distinct hospital systems, one of which was never provided to the participants. While several models had reasonable performance on hospitals they had at least partial access to, scores dropped substantially on the third, unseen hospital. Models using more sophisticated modifications of IM features saw a larger drop than those using simple binary variables or no representation of health care processes.

Objectives

In this study we seek to empirically verify and understand the effect that including IM features has on health care machine learning models. We selected 3 common outcomes of interest, mortality, length-of-stay, and sepsis, and trained models for 2 task settings. The first, shared by all outcomes, is entire record classification where the model provides a prediction at the end

of a patient's ICU stay. The second is hourly prediction of label, and only the sepsis label is used for this task.

We verify the effect of IM feature inclusion on performance, generalizability, and clinical utility of models in 3 steps. First, to get a comprehensive understanding of model performance, binary classification models for each of the outcomes were trained and evaluated using multiple metrics. Since class imbalance varies between outcomes, we could also evaluate model robustness. Second, for the sepsis outcome, since data from 2 distinct hospital systems were available, we could evaluate model generalizability and test whether that is affected by IM features. Third, again for the sepsis outcome, since labels for every hour of patient data were available, we trained a model for temporal prediction of sepsis. We evaluated this model on the hidden hospital system's data in a simulated prospective manner, in the process understanding how the models would behave in an actual clinical setting and what differences in performance can be expected by including IM features.

Finally, we hypothesized that health care processes vary across patient demographics and ICU types, which may result in varying missingness rates and patterns across subgroups. Previous works have shown how laboratory variation (and thus test ordering behavior) may vary based on these criteria; this was also seen in our data analysis [32,33]. Thus, we were motivated to see model performances for different subgroups, as well as to study the different extent to which IM features improve model performance within a subgroup. Based on our data analysis, age and ICU type subgroups were chosen. Since testing was also done on the hidden hospital, we could see how generalization on subgroups is affected by including IM. We could also verify whether models can use IM features to capture the relationship between test ordering and patient pathophysiology despite intra- and interhospital variations.

Methods

In this section we describe the datasets used for this study and the preprocessing pipeline. We also describe how outcomes of interest were defined. This is followed by an overview of the task settings and experiments with model implementation details.

Datasets

Data from the PhysioNet 2012 and 2019 Challenges were used for this study. From the PhysioNet 2012 [34] dataset (P12), we used patient records from training set A and open test set B, each consisting of data from 4000 patients collected from 4 types of ICUs. Several patient outcomes are provided of which we selected in-hospital death (mortality) and length of stay (number of days between patient's admission to the ICU and end of hospitalization, LOS). We binarized the LOS outcome setting as 3 days as a heuristic decision threshold, similar to previous studies [14]. The data consist of static patient descriptors as well as temporal variables representing patient vitals (low missingness) and values from laboratory tests conducted (high missingness). Imbalance ratios of mortality and LOS were different, at 13.9% and 6.5%, respectively, for set A and 14.2% and 7.0%, respectively, for set B. Since P12 was extracted from the MIMIC II (Multiparameter Intelligent Monitoring in Intensive Care) Clinical Database [35], the data were from one hospital system only.

The PhysioNet 2019 [29] dataset (P19) comprised patient records from 3 geographically distinct US hospital systems. A total of 40,336 patient records, 20,336 from hospital A (set A) and 20,000 from hospital B (set B), from 2 ICU types were used. Data from hospital C were not available for download. Since the challenge was aimed at model development for early prediction of sepsis, a corresponding binary label is provided for every hour of the patient's record. Labeling was done in accordance with the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) criteria [36]. It is important to note that to facilitate training models for early prediction, patients who eventually developed sepsis were labeled as such starting 6 hours before a confirmed diagnosis. More details about the definition of the sepsis label may be found in Multimedia Appendix 1. Available variables in the dataset are similar to that in P12, describing static as well as temporal patient features with varying missingness. The cohort from hospital A consisted of 8.8% of patients who developed sepsis while it was 5.7% for hospital B. Due to the cohort selection procedure followed by Reyna et al [29], few patients have sepsis from the start of ICU admission. Only 2.2% of hourly records for hospital A and 1.4% for hospital B are labeled as corresponding to sepsis. For analysis of the extent of missingness in the various datasets, please see Figures S1-S5 in Multimedia Appendix 2.

Preprocessing

Data preprocessing was done using a similar pipeline as described in multiple previous studies [13,14,37]. Data from P12 were resampled on an hourly basis, while P19 data were already resampled. While resampling, some patient records were found to have static descriptors only and others had missing outcome labels in both sets of P12. These were removed, leaving 3997 patient records in set A and 3993 in set B. Invasive and noninvasive measurements of the same variable present in P12 were averaged to form aggregate measurements. In P19, end tidal carbon dioxide was a variable observed in only hospital B, so it was removed from consideration. Static patient features describing age, gender, or ICU type identifiers were not used as inputs. This left us with 33 features in P12 and 34 in P19, which were used for model training. To deal with missing data, zero imputation was performed in both datasets, since Lipton et al [13] showed that this simple strategy proved quite effective when used to train deep learning models.

For model training and evaluation, training and testing sets were identified. Set A from both datasets was used for training while set B was shown to the model only for final evaluation. It is worth noting again that set B in P19 belonged to a distinct hospital system. Data were standardized before inputting to the model. Mean and variance from training data were used to standardize corresponding test data.

Finally, we describe the derivation of features to represent missingness. We selected the simplest representation using binary indicator variables, with a 1 used to denote variable observation and a 0 otherwise. Every feature described earlier

had a corresponding missingness indicator that was appended to the feature vector as in Lipton et al [13]. This resulted in 66 features for P12 and 68 for P19.

Modeling Methodology

Since patient pathophysiology evolves nonlinearly over time, sequential models like recurrent neural networks (RNN) are considered suitable and have often been used in previous works [38]. We used a gated RNN variant, specifically a gated recurrent unit (GRU) to model long EHR sequences [39]. A multilayer perceptron followed by a sigmoid layer were used after the GRU to output binary label probabilities.

The model was implemented in Pytorch [40] and trained using minibatch gradient descent to minimize binary cross entropy loss with Adam [41] as the optimizer. Models trained with IM augmented features are denoted by masking while those trained with patient physiological features only are denoted by no masking.

We performed 5-fold stratified cross validation for hyperparameter tuning and to prevent model overfitting. To tune hyperparameters, we performed an iterative ranging investigation to determine a suitable grid followed by a grid search [42]. Maximum averaged AUROC and utility score across all folds were chosen as the criteria for hyperparameter set selection for the retrospective and simulated prospective tasks, respectively [29]. No attempt was made to tune model architecture as our focus was not to propose a new model but to evaluate IM feature effectiveness.

Task Settings

We analyzed the effectiveness of including IM features by defining 2 tasks, (1) retrospective classification where we verify IM usefulness on model performance, calibration, and generalizability and (2) simulated prospective classification to study IM effect on model prediction trends in a temporal manner.

Retrospective Classification

In this setting, the model is trained to predict the appropriate label at the end of a patient's hospital stay. For this purpose, mortality and LOS labels were used directly from the outcomes provided in P12. For P19, a sepsis-overall label was derived from the hourly labels provided. If a patient developed sepsis at any time, their entire record was marked as positive for sepsis. The task for all 3 labels was binary classification after using the entire patient record as input. We studied the effect of IM in 2 steps, overall classification and subgroup analysis:

- To verify changes in performance on IM inclusion, the models were evaluated on all of the testing data for all datasets and labels. Multiple evaluation metrics were used to understand how IM features change performance and calibration while data from a distinct hospital were used to evaluate changes in model generalizability.
- To study extent of improvement on different patient subgroups, models were trained on all of the training data (representative of a general ICU population) and evaluated on identified subgroups made from the test set. Both datasets provided 3 general patient descriptors: age, gender, and ICU type. Visual comparison of variable observation differences between these strata was performed. Gender showed no substantial difference in variable observation. Different ICU types displayed clear differences as did age after binning into suitable intervals (Figures S6-S11 in Multimedia Appendix 2). These strata were chosen for subgroup analysis.

Simulated Prospective Classification

Only P19 was used for this task since P12 did not have hourly labels. The model was trained to predict patient probability of sepsis at every hour using the shifted labels provided in the dataset. At time *t*, information from the beginning of the patient record to *t* was used to make a prediction. This ensured prospective usefulness of the model. Since the model was trained on labels shifted by 6 hours (for septic patients), we expected the model to learn early signs of sepsis onset. The sepsis-overall label described earlier was used for cross-validation and hyperparameter tuning.

Performance Evaluation

Model discriminative ability was judged by the concordance index or AUROC. Since this is known to be an over optimistic measure for imbalanced datasets [43], we also use the precision-recall curve and average precision to evaluate predictive value [44,45]. Finally, 2 measures were used to assess model calibration: reliability plots and Brier score. The former was useful to visualize calibration changes against different levels of model uncertainty. The latter was used to quantify an averaged deviation from true probabilities and as a convenient summary of uncertainty, resolution, and reliability [46]. We also visualized the number of samples in each bin of the reliability plots by varying marker area proportional to the squared root of the bin size scaled by a constant factor. Finally, AUROC and Brier score were reported with 95% confidence intervals computed with 10,000 bootstrap replications to obtain a good estimation of model performance up to the second significant digit [47].

Results

Retrospective Classification

Overall Classification

The first 3 rows of Table 1 summarize results for the overall classification tasks. Including IM resulted in considerable improvements over using patient physiological features only for both tasks on P12 and the sepsis-overall task on P19. The extent of improvement in average precision mimicked trends of improvements in AUROC. The no masking model had an average precision of 0.493 on the P12 mortality task, and including IM features improved this to 0.511. The performance gain was more marked for the P12 LOS task, as average precision was 0.173 without and 0.368 with masking. It is worth noting that the derived LOS label in P12 had higher class imbalance than the mortality label for the same dataset. The P19 sepsis-overall task also saw an improvement in average precision where the no masking model achieved 0.537 and this was 0.547 for the masking model. Panels A and B of Figures

1-3 graphically show the receiver operating characteristic and PR curves for these tasks.

Including IM features also improved model calibration scores in all 3 cases, as seen by the Brier score (lower is better). The improved Brier scores (0.039 with IM features vs 0.045 without) for the P19 sepsis-overall task where evaluation was on a distinct hospital suggests that the model does not overfit to hospital-specific health care process variables. Examining panel C of Figures 1-3 shows the calibration plots for each task setting. The 2 models had very similar plots for the P12 mortality task. The difference was again most pronounced for the P12 LOS task, where the masking model had better calibration at higher model certainties (predicted probabilities). The masking model also showed improved calibration for the P19 sepsis-overall task seen in Figure 3C.

Table 1. Results of model discrimination and calibration for all task settings on the test data. These correspond to internal validation for PhysioNet 2012 Challenge and external for PhysioNet 2019 Challenge.

	Masking (AUROC[a]), mean (SD)	Masking (Brier), mean (SD)	No masking (AUROC), mean (SD)	No masking (Brier), mean (SD)
P12[b] mortality	0.842 (0.82-0.86)	0.093 (0.087-0.100)	0.830 (0.81-0.85)	0.095 (0.088-0.101)
P12 LOS[c]	0.814 (0.79-0.84)	0.054 (0.049-0.060)	0.737 (0.71-0.77)	0.064 (0.058-0.070)
P19[d] sepsis-overall	0.907 (0.90-0.92)	0.039 (0.036-0.041)	0.889 (0.88-0.90)	0.045 (0.043-0.048)
P19 sepsis-frequent	0.757 (0.74-0.77)	0.014 (0.013-0.014)	0.766 (0.75-0.78)	0.014 (0.013-0.015)

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

[b]P12: PhysioNet 2012 Challenge.

[c]LOS: length of stay.

[d]P19: PhysioNet 2019 Challenge.

Figure 1. Receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and calibration plot for the PhysioNet 2012 Challenge mortality classification task.

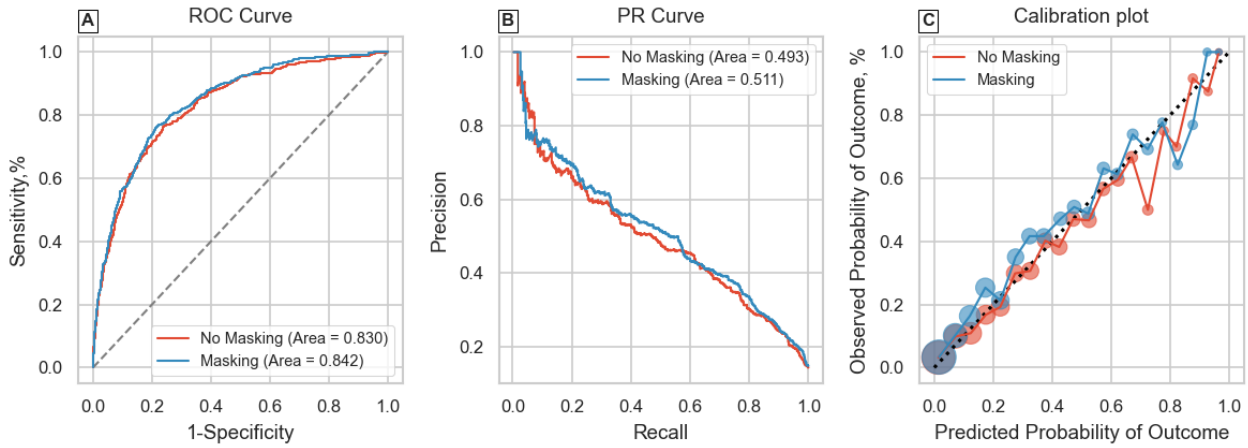


Figure 2. Receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and calibration plot for the PhysioNet 2012 Challenge length of stay classification task.

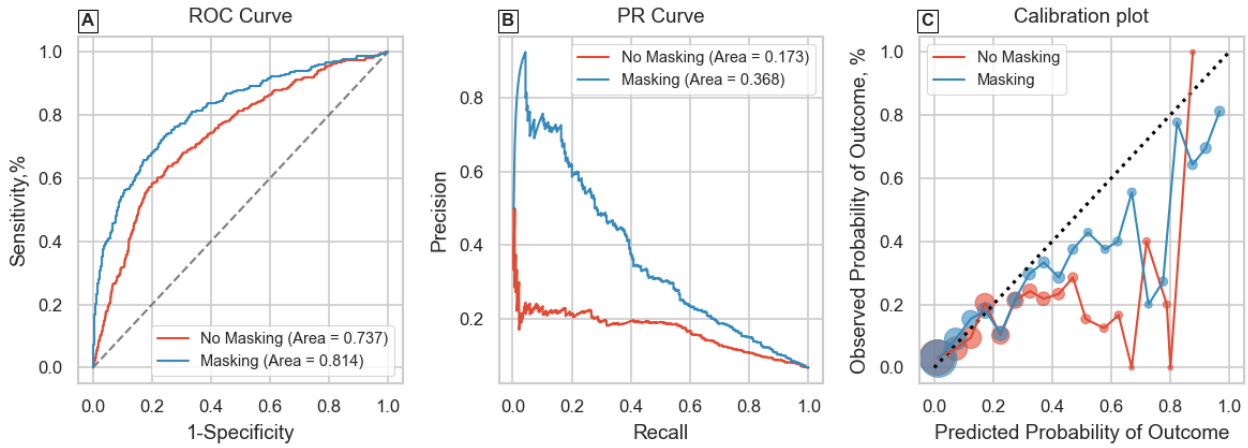
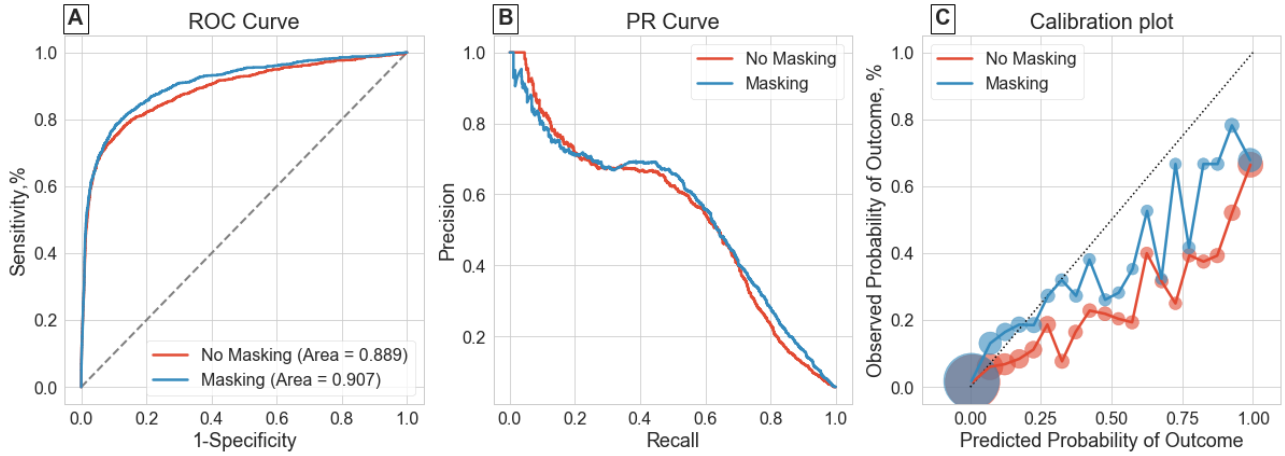


Figure 3. Receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and calibration plot for the PhysioNet 2019 Challenge sepsis-overall classification task.



Subgroup Analysis

Tables 2-4 summarize model performances on the identified subgroups for the 3 overall classification task settings. For variance estimation in results, the subgroup data were bootstrapped keeping the sample size equal to subgroup size. These results have also been visualized as bar plots in Figures S12-S14 in Multimedia Appendix 2.

For the P12 mortality task in Table 2, the no masking model outperformed the masking model for the age bins 35 years and younger and 45 to 55 years, while the masking model had better performance for all other age groups. The best AUROC over all ages was achieved by the masking model on the 35- to 45-year group, which also saw the largest improvement on including IM features (2.6%). While younger and middle-aged groups saw inconsistent performance changes on IM inclusion, older patients (older than 55 years) showed consistent improvements from 0.8% to 1.5% in all-cause mortality classification. When considering performances in different ICUs, the masking model generally had better performance except for the coronary care unit (CCU), but the difference was not substantial. The cardiac surgery recovery unit saw the highest AUROC and also the greatest improvement of 1.7% on IM inclusion.

Similar to the prominent improvements in the P12 LOS-overall classification task, the masking model considerably outperformed the no masking model for all age and ICU type subgroups. The youngest age group, 35 years and younger, saw an improvement of 15.5% in AUROC, becoming the subgroup with the best performance out of all age groups. Comparatively, the 55- to 65-year subgroup, which had the best model performance without IM, saw an improvement only of 0.7%. The cardiac surgery recovery unit again saw the largest performance gain on IM inclusion, of 13.1%, followed by the surgical ICU with 10.2% and the CCU, with a relatively small gain of 2.8%.

Finally for the P19 sepsis-overall task, the masking model again outperformed all subgroups except for the 35- to 45-year bin. Older groups (older than 55 years) generally saw a larger improvement, with the greatest increase in AUROC seen in the 65- to 75-year group, at 4%. While the surgical and medical ICUs had the same AUROC without IM, the masking model performed better on the surgical ICU.

Brier score trends generally showed similar or improved calibration on including IM features for all outcomes and subgroups. Particularly for P19 sepsis-overall, calibration improved despite external validation.

Table 2. Subgroup analysis results for the PhysioNet 2012 Challenge mortality classification task.

	#Samples	Masking (AUROC[a]), mean (SD)	Masking (Brier), mean (SD)	No masking (AUROC), mean (SD)	No masking (Brier), mean (SD)
Age strata (years)					
≤35	268	0.847 (0.74-0.93)	0.057 (0.037-0.079)	0.852 (0.75-0.94)	0.059 (0.040-0.079)
35-45	309	0.906 (0.84-0.96)	0.048 (0.031-0.066)	0.880 (0.80-0.95)	0.054 (0.037-0.072)
45-55	569	0.878 (0.82-0.93)	0.064 (0.050-0.078)	0.885 (0.83-0.93)	0.064 (0.052-0.077)
55-65	708	0.859 (0.82-0.90)	0.074 (0.060-0.090)	0.848 (0.80-0.89)	0.076 (0.063-0.090)
65-75	845	0.830 (0.79-0.87)	0.094 (0.079-0.109)	0.822 (0.78-0.86)	0.094 (0.080-0.108)
>75	1294	0.801 (0.77-0.83)	0.135 (0.121-0.149)	0.786 (0.75-0.82)	0.135 (0.123-0.149)
ICU[b] types					
Coronary care unit	587	0.806 (0.75-0.86)	0.087 (0.069-0.106)	0.807 (0.74-0.86)	0.086 (0.070-0.104)
Cardiac surgery unit	780	0.862 (0.79-0.92)	0.035 (0.025-0.046)	0.845 (0.76-0.92)	0.037 (0.028-0.048)
Surgical ICU	1192	0.852 (0.82-0.88)	0.094 (0.082-0.107)	0.843 (0.81-0.87)	0.095 (0.083-0.106)
Medical ICU	1434	0.801 (0.77-0.83)	0.128 (0.115-0.140)	0.787 (0.76-0.82)	0.129 (0.117-0.141)

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

[b]ICU: intensive care unit.

Table 3. Subgroup analysis results for the PhysioNet 2012 Challenge length of stay classification task.

	#Samples	Masking (AUROC[a]), mean (SD)	Masking (Brier), mean (SD)	No masking (AUROC), mean (SD)	No masking (Brier), mean (SD)
Age strata (years)					
≤35	268	0.862 (0.80-0.92)	0.081 (0.055-0.109)	0.707 (0.61-0.80)	0.108 (0.079-0.138)
35-45	309	0.820 (0.71-0.91)	0.060 (0.040-0.081)	0.721 (0.62-0.82)	0.079 (0.057-0.104)
45-55	569	0.800 (0.72-0.88)	0.057 (0.042-0.073)	0.712 (0.63-0.79)	0.064 (0.048-0.081)
55-65	708	0.797 (0.71-0.87)	0.045 (0.033-0.059)	0.790 (0.72-0.86)	0.054 (0.042-0.068)
65-75	845	0.803 (0.72-0.87)	0.047 (0.035-0.060)	0.712 (0.64-0.78)	0.053 (0.042-0.065)
>75	1294	0.814 (0.77-0.86)	0.056 (0.046-0.067)	0.747 (0.69-0.80)	0.062 (0.052-0.073)
ICU[b] types					
Coronary care unit	587	0.791 (0.73-0.85)	0.086 (0.068-0.105)	0.763 (0.71-0.82)	0.095 (0.078-0.112)
Cardiac surgery unit	780	0.890 (0.77-0.98)	0.013 (0.006-0.020)	0.759 (0.60-0.90)	0.018 (0.011-0.025)
Surgical ICU	1192	0.812 (0.75-0.87)	0.046 (0.036-0.056)	0.710 (0.64-0.77)	0.056 (0.036-0.056)
Medical ICU	1434	0.776 (0.73-0.82)	0.071 (0.060-0.083)	0.682 (0.63-0.73)	0.082 (0.071-0.094)

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

[b]ICU: intensive care unit.

Table 4. Subgroup analysis results for the PhysioNet 2019 Challenge sepsis-overall classification task. A total of 6095 patients did not have intensive care unit type specified, and thus, they were not considered for the corresponding analysis.

	#Samples	Masking (AUROC[a]), mean (SD)	Masking (Brier), mean (SD)	No masking (AUROC), mean (SD)	No masking (Brier), mean (SD)
Age strata (years)					
≤35	1742	0.904 (0.86-0.94)	0.037 (0.029-0.045)	0.893 (0.85-0.93)	0.044 (0.035-0.052)
35-45	1949	0.911 (0.88-0.94)	0.041 (0.033-0.049)	0.910 (0.88-0.94)	0.046 (0.038-0.055)
45-55	3334	0.920 (0.90-0.94)	0.032 (0.026-0.037)	0.900 (0.87-0.93)	0.037 (0.032-0.043)
55-65	4581	0.897 (0.87-0.92)	0.042 (0.037-0.048)	0.886 (0.86-0.91)	0.048 (0.042-0.053)
65-75	4768	0.917 (0.90-0.94)	0.039 (0.034-0.043)	0.877 (0.85-0.90)	0.049 (0.043-0.054)
>75	3626	0.896 (0.87-0.92)	0.040 (0.034-0.046)	0.888 (0.86-0.91)	0.045 (0.039-0.051)
ICU[b] types					
Medical ICU	6923	0.895 (0.88-0.91)	0.044 (0.040-0.048)	0.882 (0.86-0.90)	0.049 (0.045-0.053)
Surgical ICU	6982	0.903 (0.89-0.92)	0.041 (0.037-0.045)	0.882 (0.86-0.90)	0.050 (0.046-0.055)

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

[b]ICU: intensive care unit.

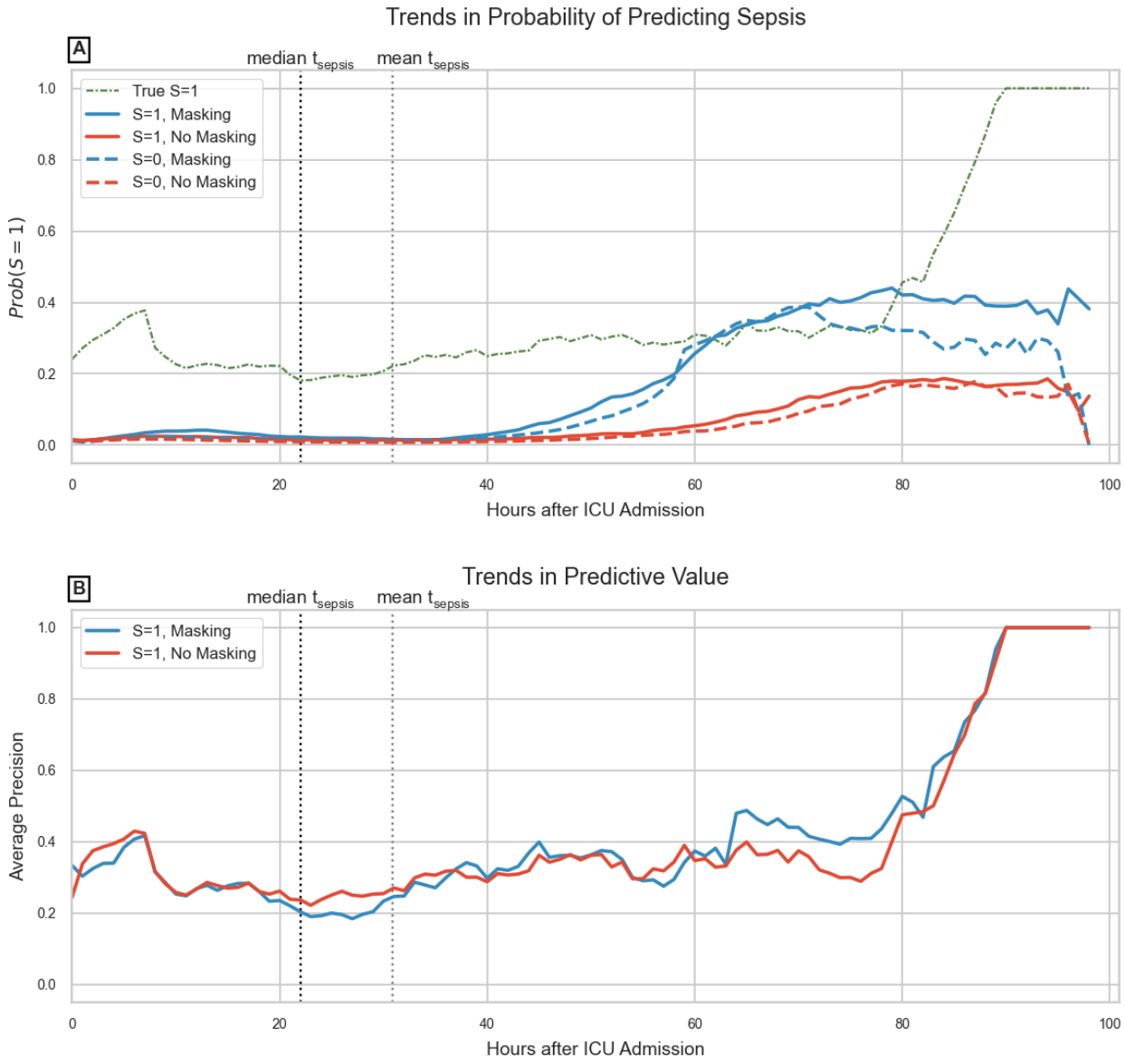
Simulated Prospective Classification

The last row of Table 1 summarizes the nontemporal evaluation for this task setting. Unlike overall classification, the no masking model outperforms the masking model while keeping almost the same calibration.

Before discussing temporal performances, it is necessary to understand the LOS distribution for each patient category. LOS averaged over the entire cohort was very similar for both P19 hospitals, at 39.77 (SD 22.55) hours and 38.23 (SD 23.27) hours for A and B, respectively. Separating the cohort into patients who eventually develop sepsis and those who don't shows that patients who develop sepsis spend a longer time in the ICU. For hospital A, septic patients spent 59.54 (SD 57.81) hours on average while nonseptic patients spent 37.87 (SD 13.92) hours. Similarly, for hospital B this was 59.22 (SD 61.90) hours for septic patients and 36.96 (SD 17.72) hours for nonseptic patients. The cohort for both hospitals consisted almost entirely of patients with sepsis after 3 days.

Temporal evaluation shown in Figure 4B displays almost equal predictive value at each hour over the first 100 hours of ICU admission, with peak predictive value achieved a little over 90 hours. This is likely due to the LOS characteristics of the datasets. Figure 4A shows how model predictions change over time for patients who eventually develop sepsis and those who don't. We observe a considerable divergence between the curves of masking and no masking models (regardless of sepsis category) a little after 2 days of ICU admission. The same plot also shows trends in the proportion of septic patients at each hour, giving an insight into the expected amount of false alarms or missed diagnoses by each model.

Figure 4. Temporal evaluation for the PhysioNet 2019 Challenge sepsis-frequent task; records corresponding to sepsis are labeled as S=1 while the remainder are S=0: (A) drop in probability of false-positive prediction (S=0) is because after 90 hours, only patients with sepsis remain in the data; (B) this cohort characteristic is learned by the model resulting in perfect predictive value after 90 hours. ICU: intensive care unit.



Discussion

Principal Findings

Results from the retrospective-overall classification shown in Table 1 were consistent with previous studies [11,13,14], confirming that including even simple representations of health care processes like binary IM features improves performance. This was further reinforced by evaluating the models on a variety of metrics summarizing predictive value and calibration. Model discrimination and predictive value were improved in all cases while keeping the same or better calibration. Results of the P19 sepsis-overall task also confirmed that model generalization in such retrospective tasks is not affected by including IM features, despite interhospital variations. Calibration plots in panel C of Figures 1-3 showed that model reliability was improved for nearly all levels of model certainty, especially for higher predicted probabilities, making the masking model more trustworthy.

Subgroup analysis helped us verify the IM inclusion effect on population subgroups and whether health care process variables encoded information about pathophysiology despite intra- and interhospital variations, justifying their use as proxy biomarkers of patient health. In the P12 mortality subgroup task (Table 2), while the masking model performed better on average in the entire test set, it failed to improve upon the no masking model for certain age groups suggesting that for younger patients, trends in physiological features alone are better predictors of in-hospital death. The masking model was also slightly outperformed by the no masking model in the CCU subgroup, which may be because CCU patients have a very specific set of complications, rendering several laboratory tests unnecessary [48]. For subgroups in P12 LOS (Table 3), however, considerable improvements in AUROC for younger age groups were observed, suggesting laboratory tests conducted were important indicators to estimate whether a patient will spend more or less than 3 days in the ICU. The CCU again saw only a slight improvement, probably due to a generally earlier

diagnosis relative to other ICUs. Overall, for both P12 outcomes, younger age groups and the cardiac surgery recovery unit had the highest AUROCs achieved by masking models.

For subgroups in the P19 sepsis-overall task (Table 4), older age groups generally saw greater benefit on IM inclusion. Sepsis is known to be associated with age, which may in turn prompt physicians to order relevant tests earlier in the patient's ICU stay [49]. The surgical ICU again saw a greater improvement in AUROC over the medical ICU, while the model had almost equal performance for both ICUs using only physiological features. This task also evaluated model performance and effect of IM features on model generalization, since the subgroups were made using data from a distinct hospital. These results suggest that, at least in retrospective task settings, health care process variables do not hinder model generalization and models trained using these variables can adequately learn the relation of IM features to the underlying condition without being affected by interhospital variations.

Calibration indicated by the Brier score showed that the model actually learns to output better probabilities on including health care process variables.

Relationship With Prior Work

Perhaps the study most similar to this work was by Sharafoddini et al [50], which examined whether missing indicator features are informative. The study performed extensive data analysis and evaluated logistic regression and tree-based models trained with and without missing indicators to assess any difference in discriminative ability. Their results demonstrated improved model performance upon IM inclusion, and feature selection methods reinforced the importance of IM variables. While this work is similarly motivated in its goal to objectively assess IM features, there are some essential differences. We focused on several outcomes of interest as opposed to mortality only, as discussed earlier. We also provided comprehensive evaluation through multiple metrics, assessing not only overall discrimination but also hourly discrimination and model calibration. Subgroup analysis and evaluation of model generalization on a distinct patient population further contribute to the novelty of this work. Previous studies did not evaluate their model's performance on ICU population subgroups, instead assuming similar performances across patients [9,13,14]. We showed that discrimination varies between strata as does the extent of improvement brought by including IM features. Finally, we used a sequential deep learning model (GRU) as opposed to the models used in Sharafoddini et al [50], since RNN variants have been popular choices to model EHR data and often use IM features to improve performances [13,14].

Temporal trends in probability of predicting sepsis shown in Figure 4A confirm previous findings by Sharafoddini et al [50] that indicators become increasingly important from the second day onward in the ICU. But this is arguably too late, since patients who eventually developed sepsis had a higher variance in LOS, many becoming septic early on in their ICU stay. While including IM features results in better model performance overall, it also falsely identifies nonsepsis patients as susceptible (false positives) in the near future, leading to several false alarms. In the PhysioNet 2019 Challenge, the utility score metric applied a minimal penalty for false positive predictions, while also leading to earlier and greater true positives, perhaps explaining the extensive use of IM features in proposed models. But alarm fatigue is a known issue in ICU early warning scores, and false positives cannot be ignored [51]. When performance on predicting the absence of sepsis (true negatives) is not considered, the net predictive value gets balanced out, as shown in Figure 4B. Also, unlike previous studies, which relied on end-of-day outcome prediction or thresholded decision outputs for evaluation, we relied exclusively on hourly probabilities and visualized its trends with time, which may be used to understand a model's clinical utility more comprehensively [27,42].

It is important to understand that IM feature effectiveness varies based on the outcome of interest, whether they are applied for retrospective or prospective tasks and even on population subgroups. With IM features now being used for a variety of tasks including classification, prediction, and even imputation, models relying on these may further propagate preexisting biases in health care processes.

Limitations

A limitation of this study was using data from the same country, in this case the United States. Practices and case-mix vary by country. Physician attitudes to uncertainty (which may influence test ordering and drug prescription) may also be affected by resource limitations and even by cultural factors [24]. This requires verifying masking model generalizability on data from different parts of the world. Efforts have been made to standardize test ordering behavior but guidelines are followed to varying extents depending on patient histories, comorbidities, and the physician in charge [26,52].

The datasets we used were observational, with no information regarding the context in which laboratory tests were ordered or which patients were transfers from other ICUs. The latter leads to the problem of lead-time bias, which may be reflected in the data as unexpected adverse outcomes for certain patients [53]. We also evaluated IM feature effectiveness on only one model type, GRU (an RNN variant). While we selected this because of its common use in prior work, different models may learn IM representations differently [38].

Critical care EHRs are also a specific subtype of general EHRs, since they consists only of inpatients with serious conditions. A more general EHR dataset that includes outpatients may result in different health care process observation patterns and reveal interesting effects on predictive models [23]. Finally, clinical best practices change over time, in turn affecting which tests are performed and how often. This is part of the larger problem of dataset shift in machine learning, and it remains to be seen how this would affect clinical models relying on health care process features.

Conclusion and Future Work

With increasing use of observational EHR data for machine learning model development, there has been an increase in the number of studies claiming clinical utility of proposed models, many relying on variables representative of health care processes. In this study, we addressed questions regarding the

effect of using health care process features on machine learning model performance and generalizability. By separating commonly used task settings into 2 subtypes, retrospective and (simulated) prospective, we made an important distinction concerning possible clinical utility of models. We framed all our results using multiple evaluation metrics while also analyzing external validation performances for all tasks by using data from a geographically distinct hospital.

This study demonstrated the usefulness of IM features in retrospective task settings on various outcome labels. Notably, we found that machine learning model generalization and calibration are not adversely affected on using health care process variables even when externally evaluated. However, the extent of improvement may depend on different patient and in-hospital factors such as age or ICU type. Our research indicated that these features provide better information for certain subgroups than others, and IM variables are better predictors of administrative outcomes like length of stay than mortality or sepsis. Results also showed that, at least for a sequential deep learning model, using simple binary missingness indicators for simulated prospective sepsis classification did not add any benefit over a model relying on patient pathological features only.

Our findings suggest that the suitability of using IM features in machine learning models may vary based on the outcome of interest, subgroup of application, task setting (retrospective or prospective), and differences in clinical practice between training data and test data. Class imbalances and nature of outcome have an intense impact on expected performance improvements on IM feature inclusion. In application, the subgroup of a patient and deviation in model performance from its expectation also need to be considered while estimating the uncertainty of a prediction. Also, while ultimately machine learning models aim to lend themselves to use as continuous monitoring bedside tools, using IM features does not seem to add any prominent improvement over not using them in that setting. Finally, using IM means using clinical practice variables in a model, so different missingness rates and missingness patterns need to be properly contextualized to understand model performance differences between train and test environments. Biased observations in one dataset (due to practice or even hospital resource variations) may have a substantial effect on model discriminations and calibration in another dataset.

There are several ways to extend this study. Future work may (1) focus on verifying model performance and generalization changes by using data from multiple countries, (2) focus on using different types of models and analyze how differently learned representations of missingness affect performance, or (3) study how health care process features may be used for multilabel classification.

Conflicts of Interest

The authors have applied for a related patent (Japanese patent application number 2019-164443).

Multimedia Appendix 1
Description of the sepsis label in PhysioNet 2019 Challenge.
[PDF File (Adobe PDF File), 46 KB - medinform_v9i12e25022_app1.pdf]

Multimedia Appendix 2
Further data and result analysis.
[PDF File (Adobe PDF File), 3113 KB - medinform_v9i12e25022_app2.pdf]

References

1. Ghassemi M, Pimentel MAF, Naumann T, Brennan T, Clifton DA, Szolovits P, et al. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data. Proc Conf AAAI Artif Intell 2015 Jan;2015:446-453 [FREE Full text] [Medline: 27182460]
2. Che Z, Kale D, Li W, Bahadori MT, Liu Y. Deep computational phenotyping. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2015:507-516. [doi: 10.1145/2783258.2783365]
3. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining 2015:1721-1730. [doi: 10.1145/2783258.2788613]
4. Hug CW, Szolovits P. ICU acuity: real-time models versus daily models. AMIA Annu Symp Proc 2009 Nov 14;2009:260-264 [FREE Full text] [Medline: 20351861]
5. 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 Dec;22(5):1589-1604. [doi: 10.1109/JBHI.2017.2767063] [Medline: 29989977]
6. Shickel B, Loftus TJ, Adhikari L, Ozrazgat-Baslanti T, Bihorac A, Rashidi P. DeepSOFA: a continuous acuity score for critically ill patients using clinically interpretable deep learning. Sci Rep 2019 Feb 12;9(1):1879 [FREE Full text] [doi: 10.1038/s41598-019-38491-0] [Medline: 30755689]

7. Lasko TA, Denny JC, Levy MA. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. PLoS One 2013;8(6):e66341 [FREE Full text] [doi: 10.1371/journal.pone.0066341] [Medline: 23826094]
8. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011 Mar;20(1):40-49 [FREE Full text] [doi: 10.1002/mpr.329] [Medline: 21499542]
9. Wei C, Dong W, Jian L, Hao Z, Lei L, Yitan L. BRITS: Bidirectional recurrent imputation for time series. Adv Neural Inf Proc Syst 2018:6775-6785 [FREE Full text]
10. Rubin DB. Inference and missing data. Biometrika 1976;63(3):581-592. [doi: 10.1093/biomet/63.3.581]
11. Lin J, Haug PJ. Exploiting missing clinical data in Bayesian network modeling for predicting medical problems. J Biomed Inform 2008 Feb;41(1):1-14 [FREE Full text] [doi: 10.1016/j.jbi.2007.06.001] [Medline: 17625974]
12. De Brouwer E, Simm J, Arany A, Moreau Y. Gru-ode-bayes: continuous modeling of sporadically-observed time series. Adv Neural Inf Proc Syst 2019:7379-7390.
13. Lipton ZC, Kale DC, Wetzel R. Modeling missing data in clinical time series with RNNs. ArXiv. Preprint posted online June 13, 2016. [FREE Full text]
14. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep 2018 Apr 17;8(1):6085 [FREE Full text] [doi: 10.1038/s41598-018-24271-9] [Medline: 29666385]
15. Øyvind Mikalsen K, Soguero-Ruiz C, Maria Bianchi F, Revhaug A, Jenssen R. Time series cluster kernels to exploit informative missingness and incomplete label information. ArXiv. Preprint posted online on July 10, 2019. [doi: 10.1016/j.patcog.2021.107896]
16. Janmajay S, Kentaro O, Raghava K, Masahiro S, Tomoko O, Noriji K. Utilizing informative missingness for early prediction of sepsis. Comput Cardiol 2019:1-4. [doi: 10.22489/cinc.2019.280]
17. Morrill J, Kormilitzin A, Nevado-Holgado A, Swaminathan S, Howison S, Lyons T. The signature-based model for early detection of sepsis from electronic health records in the intensive care unit. Comput Cardiol 2019:1. [doi: 10.22489/cinc.2019.014]
18. Zabihi M, Kiranyaz S, Gabbouj M. Sepsis prediction in intensive care unit using ensemble of xgboost models. Comput Cardiol 2019:1. [doi: 10.22489/cinc.2019.238]
19. Kristiansen IS, Hjortdahl P. The general practitioner and laboratory utilization: why does it vary? Fam Pract 1992 Mar;9(1):22-27. [doi: 10.1093/fampra/9.1.22] [Medline: 1634022]
20. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. AMIA Annu Symp Proc 2013;2013:1472-1477 [FREE Full text] [Medline: 24551421]
21. Rusanov A, Weiskopf NG, Wang S, Weng C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. BMC Med Inform Decis Mak 2014 Jun 11;14:51. [doi: 10.1186/1472-6947-14-51] [Medline: 24916006]
22. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 2018 Apr 30;361:k1479 [FREE Full text] [doi: 10.1136/bmj.k1479] [Medline: 29712648]
23. Pivovarov R, Albers DJ, Sepulveda JL, Elhadad N. Identifying and mitigating biases in EHR laboratory tests. J Biomed Inform 2014 Oct;51:24-34 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.016] [Medline: 24727481]
24. Zaat JO, van Eijk JT. General practitioners' uncertainty, risk preference, and use of laboratory tests. Med Care 1992 Sep;30(9):846-854. [doi: 10.1097/00005650-199209000-00008] [Medline: 1518316]
25. Leurquin P, Van Casteren V, De Maeseneer J. Use of blood tests in general practice: a collaborative study in eight European countries. Eurosentinel Study Group. Br J Gen Pract 1995 Jan;45(390):21-25 [FREE Full text] [Medline: 7779470]
26. Freedman DB. Towards better test utilization: strategies to improve physician ordering and their impact on patient outcomes. Electr J Int Fed Clin Chem 2015 Jan;26(1):15-30 [FREE Full text] [Medline: 27683478]
27. Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Sci Transl Med 2015 Aug 5;7(299):299ra122. [doi: 10.1126/scitranslmed.aab3719] [Medline: 26246167]
28. Futoma J, Hariharan S, Sendak M, Brajer N, Clement M, Bedoya A, et al. An improved multi-output gaussian process rnn with real-time validation for early sepsis detection. ArXiv. Preprint posted online on August 19, 2017. [FREE Full text]
29. Reyna MA, Josef CS, Jeter R, Shashikumar SP, Westover MB, Nemati S, et al. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Crit Care Med 2020 Feb;48(2):210-217 [FREE Full text] [doi: 10.1097/CCM.0000000000004145] [Medline: 31939789]
30. Vollmer M, Luz C, Sodmann P, Sinha B, Kuhn SO. Time-specific metalearners for the early prediction of sepsis. Comput Cardiol 2019:1-4. [doi: 10.22489/cinc.2019.029]
31. Chang Y, Rubin J, Boverman G. A multi-task imputation and classification neural architecture for early prediction of sepsis from multivariate clinical time series. Comput Cardiol 2019:1. [doi: 10.22489/cinc.2019.110]
32. Taylor JB. Relationships among patient age, diagnosis, hospital type, and clinical laboratory utilization. Clin Lab Sci 2005;18(1):8-15. [Medline: 15747782]
33. Gershengorn HB, Garland A, Gong MN. Patterns of daily costs differ for medical and surgical intensive care unit patients. Ann Am Thorac Soc 2015 Dec;12(12):1831-1836. [doi: 10.1513/AnnalsATS.201506-366BC] [Medline: 26393984]

34. Silva I, Moody G, Scott DJ, Celi LA, Mark RG. Predicting in-hospital mortality of ICU patients: the PhysioNet/Computing in Cardiology Challenge 2012. Comput Cardiol 2012;39:245-248 [FREE Full text] [Medline: 24678516]
35. Saeed M, Villarroel M, Reisner AT, Clifford G, Lehman L, Moody G, et al. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. Crit Care Med 2011 May;39(5):952-960 [FREE Full text] [doi: 10.1097/CCM.0b013e31820a92c6] [Medline: 21283005]
36. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016 Feb 23;315(8):801-810. [doi: 10.1001/jama.2016.0287] [Medline: 26903338]
37. Lipton Z, Kale D, Elkan C, Wetzel R. Learning to diagnose with lstm recurrent neural networks. ArXiv. Preprint posted online on November 11, 2015. [FREE Full text]
38. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc 2018 Oct 01;25(10):1419-1428 [FREE Full text] [doi: 10.1093/jamia/ocy068] [Medline: 29893864]
39. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D. Learning phrase representations using rnn encoder-decoder for statistical machine translation. ArXiv. Preprint posted online on June 3, 2014. [doi: 10.3115/v1/d14-1179]
40. Paszke A, Gross S, Massa F, Lerer A. Pytorch: an imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, editors. Adv Neural Inf Proc Syst. Red Hook: Curran Associates, Inc; 2019:8024-8034.
41. Kingma D, Ba J. Adam: a method for stochastic optimization. ArXiv. Preprint posted online on December 22, 2014. [FREE Full text]
42. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLoS One 2018;13(11):e0206862. [doi: 10.1371/journal.pone.0206862] [Medline: 30427913]
43. Maslove DM. With severity scores updated on the hour, data science inches closer to the bedside. Crit Care Med 2018 Mar;46(3):480-481. [doi: 10.1097/CCM.0000000000002945] [Medline: 29474330]
44. Leisman DE. Rare events in the ICU: an emerging challenge in classification and prediction. Crit Care Med 2018 Mar;46(3):418-424. [doi: 10.1097/CCM.0000000000002943] [Medline: 29474323]
45. Ozenne B, Subtil F, Maucort-Boulch D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. J Clin Epidemiol 2015 Aug;68(8):855-859. [doi: 10.1016/j.jclinepi.2015.02.010] [Medline: 25881487]
46. Ovadia Y, Fertig E, Ren J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. ArXiv. Preprint posted online on June 6, 2019. [FREE Full text]
47. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12:77 [FREE Full text] [doi: 10.1186/1471-2105-12-77] [Medline: 21414208]
48. Mehta NJ, Khan IA. Cardiology's 10 greatest discoveries of the 20th century. Tex Heart Inst J 2002;29(3):164-171 [FREE Full text] [Medline: 12224718]
49. Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Crit Care Med 2001 Jul;29(7):1303-1310. [doi: 10.1097/00003246-200107000-00002] [Medline: 11445675]
50. Sharafoddini A, Dubin JA, Maslove DM, Lee J. A new insight into missing data in intensive care unit patient profiles: observational study. JMIR Med Inform 2019 Jan 08;7(1):e11605 [FREE Full text] [doi: 10.2196/11605] [Medline: 30622091]
51. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine learning and decision support in critical care. Proc IEEE Inst Electr Electron Eng 2016 Feb;104(2):444-466 [FREE Full text] [doi: 10.1109/JPROC.2015.2501978] [Medline: 27765959]
52. Solomon DH, Hashimoto H, Daltroy L, Liang MH. Techniques to improve physicians' use of diagnostic tests: a new conceptual framework. JAMA 1998 Dec 16;280(23):2020-2027. [doi: 10.1001/jama.280.23.2020] [Medline: 9863854]
53. Dragsted L, Jörgensen J, Jensen NH, Bönsing E, Jacobsen E, Knaus WA, et al. Interhospital comparisons of patient outcome from intensive care: importance of lead-time bias. Crit Care Med 1989 May;17(5):418-422. [doi: 10.1097/00003246-198905000-00008] [Medline: 2707011]

Abbreviations

AUROC: area under the curve of the receiver operating characteristic
CCU: coronary care unit
EHR: electronic health record
GRU: gated recurrent unit
ICU: intensive care unit
IM: informative missingness

LOS: length of stay
MIMIC II: Multiparameter Intelligent Monitoring in Intensive Care
P12: PhysioNet 2012 Challenge dataset
P19: PhysioNet 2019 Challenge dataset
RNN: recurrent neural network
Sepsis-3: Third International Consensus Definitions for Sepsis and Septic Shock

Edited by C Lovis, J Hefner; submitted 14.10.20; peer-reviewed by D Maslove, Z Che, G Weber; comments to author 22.11.20; revised version received 17.02.21; accepted 02.09.21; published 08.12.21.

Please cite as:
Singh J, Sato M, Ohkuma T
On Missingness Features in Machine Learning Models for Critical Care: Observational Study
JMIR Med Inform 2021;9(12):e25022
URL: https://medinform.jmir.org/2021/12/e25022
doi:10.2196/25022
PMID:34889756

Original Paper

Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study

Qingyu Chen[1], PhD; Alex Rankine[1,2]; Yifan Peng[1,3], PhD; Elaheh Aghaarabi[1,4], MSc; Zhiyong Lu[1], PhD

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, United States

[2]Harvard College, Cambridge, MA, United States

[3]Weill Cornell Medicine, New York, NY, United States

[4]Towson University, Towson, MD, United States

Corresponding Author:
Zhiyong Lu, PhD
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health
8600 Rockville Pike
Bethesda, MD, 20894
United States
Phone: 1 301 594 7089
Email: luzh@ncbi.nlm.nih.gov

Abstract

Background: Semantic textual similarity (STS) measures the degree of relatedness between sentence pairs. The Open Health Natural Language Processing (OHNLP) Consortium released an expertly annotated STS data set and called for the National Natural Language Processing Clinical Challenges. This work describes our entry, an ensemble model that leverages a range of deep learning (DL) models. Our team from the National Library of Medicine obtained a Pearson correlation of 0.8967 in an official test set during 2019 National Natural Language Processing Clinical Challenges/Open Health Natural Language Processing shared task and achieved a second rank.

Objective: Although our models strongly correlate with manual annotations, annotator-level correlation was only moderate (weighted Cohen κ=0.60). We are cautious of the potential use of DL models in production systems and argue that it is more critical to evaluate the models in-depth, especially those with extremely high correlations. In this study, we benchmark the effectiveness and efficiency of top-ranked DL models. We quantify their robustness and inference times to validate their usefulness in real-time applications.

Methods: We benchmarked five DL models, which are the top-ranked systems for STS tasks: Convolutional Neural Network, BioSentVec, BioBERT, BlueBERT, and ClinicalBERT. We evaluated a random forest model as an additional baseline. For each model, we repeated the experiment 10 times, using the official training and testing sets. We reported 95% CI of the Wilcoxon rank-sum test on the average Pearson correlation (official evaluation metric) and running time. We further evaluated Spearman correlation, R^2, and mean squared error as additional measures.

Results: Using only the official training set, all models obtained highly effective results. BioSentVec and BioBERT achieved the highest average Pearson correlations (0.8497 and 0.8481, respectively). BioSentVec also had the highest results in 3 of 4 effectiveness measures, followed by BioBERT. However, their robustness to sentence pairs of different similarity levels varies significantly. A particular observation is that BERT models made the most errors (a mean squared error of over 2.5) on highly similar sentence pairs. They cannot capture highly similar sentence pairs effectively when they have different negation terms or word orders. In addition, time efficiency is dramatically different from the effectiveness results. On average, the BERT models were approximately 20 times and 50 times slower than the Convolutional Neural Network and BioSentVec models, respectively. This results in challenges for real-time applications.

Conclusions: Despite the excitement of further improving Pearson correlations in this data set, our results highlight that evaluations of the effectiveness and efficiency of STS models are critical. In future, we suggest more evaluations on the generalization capability and user-level testing of the models. We call for community efforts to create more biomedical and clinical STS data sets from different perspectives to reflect the multifaceted notion of sentence-relatedness.

(JMIR Med Inform 2021;9(12):e27386) doi:10.2196/27386

KEYWORDS

semantic textual similarity; deep learning; biomedical and clinical text mining; word embeddings; sentence embeddings; transformers

Introduction

Background

Semantic textual similarity (STS), a measure of the degree of relatedness between sentence pairs, is an important text-mining research topic [1]. STS has been widely used in biomedical and clinical domains, including information retrieval (finding relevant sentences or passages [2]), biocuration (finding key sentences for evidence attribution [3]), and question answering (finding answer-snippet candidates [4]). Despite its importance, expertly annotated STS data sets are lacking in the biomedical and clinical domains. For example, STS-related data sets in the general domain have been developed for nearly a decade, with almost 30,000 annotated sentence pairs in total [5], whereas similar data sets in the biomedical and clinical domains had only hundreds of pairs in total before 2018 [6]. The organizers of the Open Health Natural Language Processing (OHNLP) Consortium have dedicated efforts to expanding such data sets and establishing STS open challenges in the clinical domain since 2018. MEDSTS [7], consisting of 1068 curated sentence pairs, was used in the BioCreative/OHNLP challenge task in 2018 [8]. In 2019, over 1000 curated sentence pairs were added to the MEDSTS, renamed ClinicalSTS [9], which was used in the National Natural Language Processing Clinical Challenges (n2c2)/OHNLP. This work is a poststudy of the n2c2/OHNLP challenge.

Overall, 33 teams submitted 87 models to the n2c2/OHNLP challenge task; Pearson correlation was used as the evaluation measure, ranging from −1 (strong negative relationship) to 1 (strong positive relationship). Our National Library of Medicine and National Center for Biotechnology Information team developed an ensemble model by leveraging a range of deep learning models from 3 categories: word embedding based, sentence embedding based, and transformer based (which is described in the following sections). This model achieved a Pearson correlation of 0.8967 in the official test set, ranking second among all of the teams (P=.88 compared with the first rank, with a Pearson correlation of 0.9010). The top 10 best team submissions demonstrated relatively close performances with Pearson correlations of 0.85 to 0.90. According to the organizer's overview, most of the top systems used deep learning models [9].

A Pearson correlation of approximately 0.9 suggests that the model's predictions have a very strong correlation with gold standard annotations [10]. Such results might give the impression that deep learning models have already solved STS in the clinical domain. Nevertheless, the human-level correlation in this data set is significantly lower; for example, the agreement between 2 annotators in ClinicalSTS had a weighted Cohen κ of 0.6 [9], suggesting that only a moderate level of correlation was achieved by human experts [10]. Therefore, we urge caution with regard to the extremely high correlation achieved by the models (which might be potentially due to overfitting) and argue that it is critical to understand how these models perform in reality rather than further improve the performance in this data set. Therefore, in this postchallenge study, we aim to analyze the effectiveness and efficiency of 5 deep learning models in depth:

- For effectiveness, we investigate how a single deep learning model performs in this specific data set and further analyze the robustness of models in sentence pairs of different degrees of similarity.
- For efficiency, we measure the inference time taken by the deep learning models in the testing set. This is an important indicator of whether these models can be used in real-time applications, such as sentence search engines. To the best of our knowledge, few studies on STS in the biomedical and clinical domains have considered model efficiency. However, given that models have already achieved a Pearson correlation of approximately 0.90, measuring efficiency is arguably more important, as it quantifies whether these models could be used in production.

The principal findings are 2-fold. First, a single deep learning model trained directly on the official training set only (ie, without more advanced techniques, such as multitask learning and transfer learning) could already achieve a maximum Pearson correlation of 0.87; however, the training set's robustness to sentence pairs of different similarity levels differs significantly. A particular observation is that BERT models made the most errors (a mean squared error of over 2.5) on highly similar sentence pairs (similarity no less than 4). BERT models cannot capture highly similar sentence pairs effectively when they have different negation terms or word orders. Second, although the deep learning models achieved relatively close Pearson correlations (from 0.82 to 0.87; single models), the time efficiency differed dramatically. For example, the difference in Pearson correlations of BERT and sentence embedding models was within 0.002, but the inference time of BERT models was approximately 50 times greater than that of sentence embedding models. This brings practical challenges to using BERT models in real-time applications, especially without the availability of graphics processing units (GPUs). Furthermore, although there has been a tremendous effort to make ClinicalSTS available to the community, their source corpora inevitably limit the diversity of sentence pairs and annotation inconsistencies. Thus, we call for community efforts to create more STS data sets from different perspectives to reflect the multifaceted notion of sentence relatedness; this, in turn, will further improve the generalization performance of deep learning models.

Here, we introduce popular deep learning STS methods that have been used in the biomedical and clinical domains. The methods are broadly categorized in terms of the language models applied: word embeddings, sentence embeddings, and transformers.

Word Embedding–Based Models

Word embeddings are relatively early language models that significantly change how text is modeled. The semantic of each word is represented in a high-dimensional vector trained on large-scale corpora in an unsupervised manner. Primary word embedding methods include (1) word2vec, based on local contexts, such as using a word as input to predict its nearby words [11]; (2) Glove, based on global co-occurrence statistics [12]; and (3) fastText, which extends word2vec by adding word n-grams [13]. Many word embedding variations (eg, pretrained in the biomedical or clinical corpora, integrated with entities, and adopted retrofitting methods) are publicly available [14-16]. First, word embedding–based STS models use these embeddings to obtain vector representations of the words in sentence pairs and then use either Convolutional Neural Networks (CNNs) or recurrent neural networks to process (typically to obtain spatial or semantic patterns), followed by fully-connected layers to make predictions [16].

Sentence Embedding–Based Models

Sentence embeddings extend word embeddings by modeling sentence-level representations. The primary methods include (1) Doc2vec, similar to word2vec, using a word as input and predicting the paragraph rather than nearby words [17]; (2) FastSent, using a sentence as input and predicting the adjacent sentences [18]; and (3) SentVec, which extends word2vec and fastText by using both words (and their n-grams) and the associated sentences as inputs for training [19]. Compared with word embedding–based models, sentence embedding–based STS models are simpler: first, they use sentence embeddings to obtain sentence vectors and then use fully-connected layers for predictions [20].

Transformer-Based Models

Transformers are recent language models that revolutionize text representation methods. Using a self-attention mechanism, this model can capture long-range dependencies [21]. Transformer-based language models, such as BERT [22] and GPT [23], have replaced recurrent neural networks for many text-based applications. To date, many transformers pretrained in the general or biomedical and clinical domains are publicly available [24-27]. Similar to sentence embedding–based models, transformer-based STS models directly use transformers to obtain sentence representations and then use fully-connected layers for predictions [22].

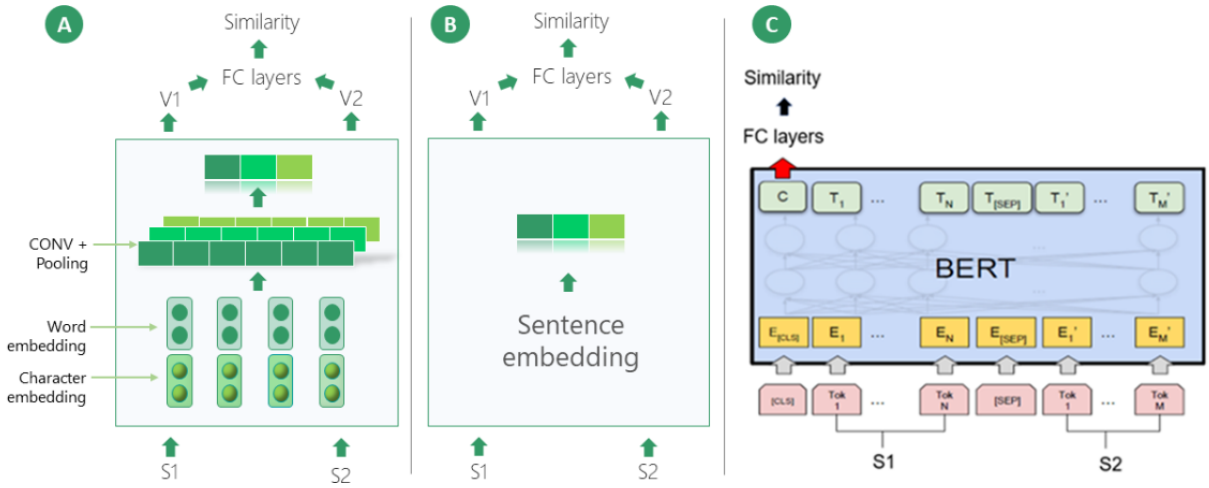
Methods

Sentence Similarity Models

Overview

Five deep learning STS models from the 3 categories above were benchmarked: the Convolutional Neural Network (CNN) model [28] (from the word embedding–based category), the sentence embedding model, using BioSentVec [29] (from the sentence embedding–based category), and transformer models (from the transformer-based category), using BioBERT [24], BlueBERT [25], and ClinicalBERT [26]. We chose these models because they achieved top-ranked performance in STS-based tasks [5,8,9]. The general architecture is shown in Figure 1, and the descriptions are as follows.

Figure 1. Model architecture overview. (A), (B), and (C) demonstrate the architecture of the Convolutional Neural Network (CNN), BioSentVec, and Bidirectional Encoder Representations from Transformers models, respectively. Details are provided in the Methods section. BERT: Bidirectional Encoder Representations from Transformers; CONV: convolutional layer; FC: fully-connected layer.



Word Embedding–Based Model (CNN Model)

We adapted the CNN model from a study by Shao [28] a top-ranked system in SemEval-2017 Task 1. The CNN model transforms the input sentence pair into vectors and learns the similarities between the corresponding vectors. The backbone is a Siamese neural network, whereby the model weights are shared when processing the 2 input sentences. The model consists of 3 layers (shown in Figure 1A). The first embedding layer consists of word and character embeddings. It is used to transform the raw text into a 2D semantic vector space. In this study, we evaluated several word embeddings. We found that word embeddings pretrained in the biomedical and clinical domains did not have additional advantages in this specific data set. This observation is consistent with the previous word embedding evaluation using the same source data set [15]. Therefore, we used Glove pretrained in the general domain for the following experiments. The second layer consists of convolutional and max-pooling layers to extract special information from the embeddings. Therefore, the 2D semantic

vector space is transformed into a 1D vector to represent the semantics of a sentence. The third layer provides a calculation of the absolute difference and dot product between the vectors of the 2 sentences. This is followed by the fully-connected layers to produce the final similarity prediction.

Sentence Embedding Model (BioSentVec Model)

We used the model from [29], which achieved the highest performance on MEDSTS for the post–BioCreative/OHNLP challenge task [20]. The model structure is similar to the CNN model above, as shown in Figure 1B. The primary difference is that this model uses BioSentVec to directly produce the sentence vectors. Therefore, there are no convolutional or pooling layers.

Transformer-Based Model (BioBERT, BlueBERT, and ClinicalBERT Models)

This model structure is illustrated in Figure 1C. First, the sentences were concatenated as one input (as recommended by the authors of BERT [22]), followed by a BERT module and fully-connected layers. We benchmarked 3 different BERT modules: (1) BioBERT [24], pretrained on PubMed abstracts and PubMed Central full-text articles; (2) BlueBERT [25], pretrained on PubMed abstracts and Medical Information Mart for Intensive Care-III clinical notes; and (3) ClinicalBERT [26], pretrained on clinical notes using the weights from BioBERT.

Additional Machine Learning Baseline Model (Random Forest)

Although the top-performing submissions used deep learning–based models [30], it is also critical to compare with traditional machine learning–based models to better understand the effectiveness and efficiency of deep learning–based models. Therefore, we evaluated the performance of a classic machine learning model as an additional baseline. Specifically, we adapted the random forest model, which achieved the best performance out of 13 submissions in the 2018 BioCreative/OHNLP challenge task [20,30]. This model uses manually engineered features in 5 dimensions to capture sentence similarity: token-based, character-based, sequence-based, semantic-based, and entity-based. We performed feature selection based on the performance of the validation set and ultimately selected 13 features.

Data Set, Evaluation Metric, and Hyperparameter Tuning

The details of the data set are presented in the data description studies [7,9]. In short, the data set consists of 2054 sentence pairs, with the similarity annotated on a scale of 0 to 5: (1) 0, if the 2 sentences are entirely dissimilar; (2) 1, if the 2 sentences are dissimilar but have the same topic; (3) 2, if the 2 sentences are not equivalent but share some details; (4) 3, if the 2 sentences are roughly equivalent but some important information is different; (5) 4, if the 2 sentences are mostly equivalent and only minor details differ; and (6) 5, if the 2 sentences are semantically equivalent [7]. The data set was annotated by 2 medical experts, with a weighted Cohen κ of 0.60 as the interannotator agreement measure [9].

The training and testing sets were officially released by the task organizers and consisted of 1642 and 429 sentence pairs, respectively. We randomly sampled approximately 20% of the sentence pairs (329 pairs) from the training set as the validation set. The Pearson correlation coefficient was used as the official evaluation metric.

Given that the models have different architectures and hyperparameters, we performed hyperparameter tuning for the CNN, BioSentVec, and BERT models separately, rather than using the same values. The values of the hyperparameters are listed in Table 1.

Table 1. Hyperparameters of the sentence similarity models. Common hyperparameters are shared among all of the models. In contrast, model-specific hyperparameters are only for specific models.

Hyperparameters	CNN[a]	BioSentVec	BERT[b] variation
Common hyperparameters			
FC[c] layers	128	512, 256, 128, 32	128, 32
Dropout	0.5	0.5	0.5
Optimizer	Adam	SGD[d]	AdamWarmup
Learning rate	1e-3	5e-3	2e-5
Batch size	64	16	32
Specific hyperparameters			
Maximum length	170	N/A[e]	128
Conv[f]	1800	N/A	N/A
Pooling	Maximum	N/A	Maximum

[a]CNN: Convolutional Neural Network.

[b]BERT: Bidirectional Encoder Representations from Transformers.

[c]FC: fully-connected.

[d]SGD: stochastic gradient descent

[e]N/A: not applicable.

[f]Conv: convolutional layers.

Evaluation Methods

We measured the Pearson correlation (for effectiveness) and the running time in seconds (for efficiency) on the testing set. To compare the 5 models quantitatively, we repeated the experiments 10 times on the same training, validation, and testing sets and reported the results of Wilcoxon rank-sum test on the average Pearson correlation and running time at 95% CI. We chose the same evaluation metric and statistical test as the task organizers for consistency [9]. We further evaluated the Spearman correlation, R^2, and mean square error as additional metrics for effectiveness.

In practice, the running time can be significantly affected by the computing environment rather than the model architecture. For instance, GPUs could significantly boost the inference time; however, many sentence search servers (especially research tools) may not have GPUs available. Different multi-processing methods may have an impact on the running time as well. For a fair comparison, we used a single processor on the central processing unit for model inference on the testing set and tracked the running time accordingly.

Results

Effectiveness and Efficiency Results

Table 2 presents the effectiveness and efficiency results. All 5 deep learning models had reasonable and very close effectiveness results for this data set. The difference between the average Pearson correlation was within 3%. The BioSentVec model achieved the highest Pearson correlation (0.8497), followed by BioBERT (0.8481; *P*=.74). The deep learning models had approximately 15% higher Pearson correlation than the baseline random forest model. In addition, the results demonstrate that a single deep learning model can achieve a maximum Pearson correlation score of 0.87. We further developed a model by averaging the predictions of the 4 best models. The ensemble model further improved the score by close to 0.90. This observation is consistent with our submission results. Table 3 provides additional effectiveness measures. BioSentVec consistently showed the highest performance in 3 out of 4 metrics, followed by BioBERT.

Table 2. Effectiveness and efficiency results for the official test set. The models are ranked by the mean effectiveness results in descending order. The *P* value of the Wilcoxon rank-sum test at a 95% CI is shown for each model compared with the model with the highest effectiveness or efficiency results. The results of the ensemble model also are provided; however, this study focuses on single models in terms of, for example, their robustness to sentence pairs of different similarity levels and their inference time for production purposes.

Model	Effectiveness (Pearson correlation)			Efficiency (seconds)		
	Values, mean (SD)	*P* value	Maximum effectiveness	Values, mean (SD)	*P* value	Lowest efficiency
Five benchmarking models						
BioSentVec	0.8497 (0.0099)	N/A[a]	0.8654	1.48 (0.23)	N/A	1.96
BioBERT	0.8481 (0.0122)	.74	0.8698	85.05 (4.93)	<.001	95.66
ClinicalBERT	0.8442 (0.0161)	.39	0.8677	85.20 (4.74)	<.001	95.21
BlueBERT	0.8320 (0.0232)	.02	0.8613	84.81 (1.63)	<.001	88.22
CNN[b]	0.8224 (0.0043)	<.001	0.8307	4.35 (0.27)	<.001	4.97
Additional machine learning baseline model						
Random forest	0.6848 (0.0022)	N/A	N/A	0.03 (0.00)	.99	0.03
Ensembled model						
Ensemble model	0.8782	N/A	0.8940	N/A	N/A	N/A

[a]N/A: not applicable.

[b]CNN: Convolutional Neural Network.

Table 3. Additional effectiveness results of individual models. The models are ranked by the Pearson correlation coefficient in descending order.

Model	Values, mean (SD)			
	Pearson correlation	Spearman correlation	R^2[a]	MSE[b]
Five benchmarking models				
BioSentVec	0.8497 (0.0099)	0.7708 (0.0073)	0.6705 (0.0325)	0.8709 (0.0434)
BioBERT	0.8481 (0.0122)	0.7951 (0.0100)	0.6636 (0.0275)	0.8803 (0.0362)
ClinicalBERT	0.8442 (0.0161)	0.8066 (0.0149)	0.6357 (0.0391)	0.9155 (0.0502)
BlueBERT	0.8320 (0.0232)	0.7701 (0.0244)	0.6520 (0.0544)	0.8935 (0.0670)
CNN[c]	0.8224 (0.0043)	0.7674 (0.0087)	0.6136 (0.0436)	0.9428 (0.0519)
Additional machine learning baseline model				
Random forest	0.6848 (0.0022)	0.6572 (0.0027)	0.4154 (0.0025)	1.1614 (0.0025)

[a]R^2: coefficient of determination.

[b]MSE: mean square error.

[c]CNN: Convolutional Neural Network.

In contrast to the effectiveness results, the efficiency results differed dramatically among the models. As shown in Table 1, it took about 1.5 seconds, on average, for the BioSentVec model to predict the similarities of 429 sentence pairs in the testing set; the counterpart of the CNN model took about 4.5 seconds, on average. In contrast, all BERT models require more than 80 seconds, on average, for inference.

Error Analysis

We further analyzed the common errors made by the models. Figure 2 shows the quantitative evaluations. We categorized the sentences into 5 groups based on the annotation guidelines and measured the MSE between the gold standard and predictions. Note that we did not use Pearson correlations as they are heavily influenced by the limited number of instances in small categories [20]. MSE is thus used as an alternative metric, which has also been used as a loss function for many deep learning models for regression-based applications.

Figure 2. Mean squared error (MSE) of the models for each similarity range. Each category shows the number of sentence pairs and associated MSE of the models. The overall MSE (median, SD) are also provided in the legend. CNN: Convolutional Neural Network.

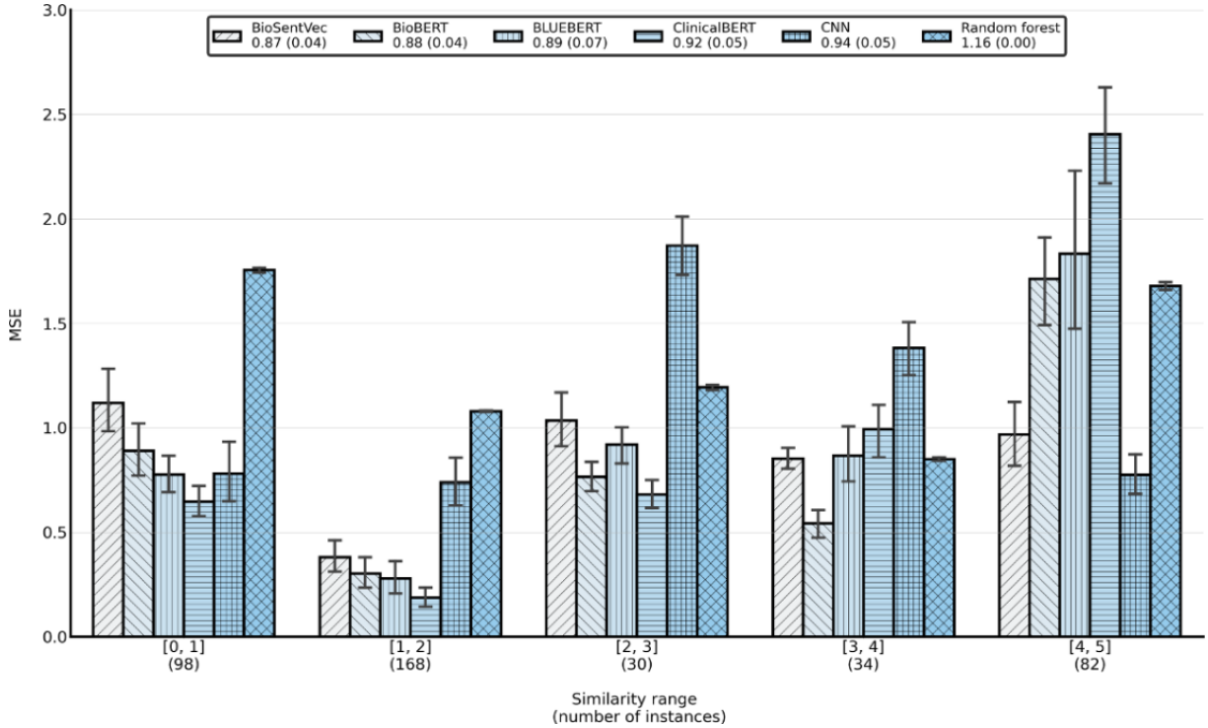


Figure 2 shows 2 primary observations. First, the random forest model had the highest MSE for the pairs with similarity scores between 0 and 1; the error rate was almost twice that of the deep learning models. In contrast, the MSEs of the random forest in other similarity categories were much smaller. This suggests that the random forest model may not effectively identify sentence pairs of low similarity. We manually examined the sentence pairs of low similarity and provided representative examples where the random forest model had a larger MSE than the other models, along with the predictions of BioBERT and BioSentVec for comparison (Table 4). The errors shared consistent patterns where (1) the sentence structure was similar (eg, both started with "The patient..."), (2) the pairs shared many common or similar words (eg, case 4 shares "examined and normal"), and (3) the semantics of the pairs were rather different. In such cases, the random forest model failed to capture the semantics at the sentence level. In addition, cases 1-3 had the gold standard annotation score of 0, whereas the similar case 5 had the counterpart of 1. One may argue that the drugs in case 5 are rather different, and the procedure was independent and could have a score of 0; alternatively, given the score of case 5, cases 1-3 could arguably have the same score as well because they were all related to patient status (similarly, both BioSentVec and BioBERT provided consistent scores on these cases). This is also consistent with the findings of the task organizers [9], which demonstrated that annotating the sentence similarity is a challenging task as relatedness is context-dependent.

Table 4. Qualitative examples with a relatively large mean squared error for the random forest model for sentence pair scores from 0.0 to 1.0.

Case	Sentence pairs	Gold standard	Random forest	BioBERT	BioSentVec
1	• The patient tolerated the procedure well and was transferred to the recovery room in stable condition. • The patient was transferred to the patient appointment coordinator for an appointment to be scheduled within the timeframe advised.	0.0	2.5	0.5	1.2
2	• Patient to call to schedule additional treatment sessions as needed otherwise patient dismissed from therapy. • Patient tolerated session without adverse reactions to therapy.	0.0	3.4	1.4	1.4
3	• Patient was agreeable to speaking with social work. • Patient was able to teach back concepts discussed.	0.0	2.0	1.7	1.7
4	• Left upper extremity: Inspection, palpation examined and normal. • Abdomen: Liver and spleen, bowel sounds examined and normal.	0.5	2.4	2.1	1.1
5	• glucosamine capsule 1 capsule by mouth one time daily. • Claritin tablet 1 tablet by mouth one time daily.	1.0	2.6	1.7	1.5

Second, all the deep learning models, except the CNN model, showed reasonable performance for the pairs with similarity scores between 1 and 4. The MSE was mainly within 1, suggesting that the predictions were likely in the same category as the gold standard. However, the BERT models had a much higher MSE for the pairs with scores from 4 to 5. For example, ClinicalBERT had an MSE of over 2.5, whereas the counterparts of both CNN and BioSentVec were lower than 1. Similarly, the variance of BERT models on sentence pairs with similarity scores from 4 to 5 was also larger than that of the other models. Table 5 shows the representative sentence pairs for which ClinicalBERT had a larger MSE than the other models, along with the predictions of BioBERT and BioSentVec for comparison. The examples indicated that ClinicalBERT could not capture highly similar sentence pairs when there are different negation terms (eg, case 1) or when the word order is switched (eg, case 2) as compared with BioBERT and BioSentVec. Similarly, interannotator consistency may also have an impact on MSE. For example, sentence pairs from cases 4 and 5 arguably belong to the same category, as the pairs share the majority of information, except for minor differences.

Table 5. Qualitative examples with a relatively large mean squared error for Bidirectional Encoder Representations from Transformers models for sentence pair scores from 4.0 to 5.0.

Case	Sentence pairs	Gold standard	ClinicalBERT	BioBERT	BioSentVec
1	• Heart: S1/S2 regular rate and rhythm, without murmurs, gallops, or rubs • Heart: S1, S2, regular rate and rhythm, no abnormal heart sounds or murmur	5.0	2.5	3.4	3.9
2	• He denies chest pain or shortness of breath • He denies shortness of breath or chest pain	5.0	2.3	3.3	3.9
3	• This patient benefits from skilled occupational and/or physical therapy to improve participation in daily occupations • Medical necessity: the patient would benefit from skilled physical therapy interventions to be able to return to work and engage in self-care activities	4.0	2.4	2.2	2.5
4	• All questions were answered to the parent's satisfaction • All questions were answered and consent was given to proceed	4.0	2.8	2.6	3.7
5	• The patient understands and is happy with the plan • The patient verbalized understanding and wishes to proceed	5.0	3.0	2.9	3.6

Discussion

Principal Findings

This study has 2 primary findings. First, the effectiveness of deep learning models on this data set is high (all 5 models have a Pearson correlation of over 0.8, which is approximately 15% higher than that of the traditional machine learning model) and relatively close (the Pearson correlation difference is within 0.03 among the models), but their efficiency is significantly different. BERT models are, on average, 20-50 times slower than the CNN and BioSentVec models, respectively.

The dramatically different efficiency results lead to the concern of using STS models in real-world applications in the biomedical and clinical domains. To demonstrate this, we further quantified the number of sentence pairs that could be computed in real-time based on the sentence search pipeline in LitSense [2]. LitSense is a web server for searching for relevant sentences from approximately 30 million PubMed abstracts and approximately 3 million PubMed Central full-text articles. To find relevant sentences for a query, it uses the standard BM25 to retrieve top candidates and then reranks the candidates using deep learning models. The rerank stage in LitSense is allocated for 300 ms based on evaluations of the developers. Using 300 ms as the threshold, BERT models can rerank only 2 pairs in real-time, whereas the CNN and BioSentVec models can rerank approximately 30 and 87 pairs, respectively. It should be noted that the results here are for demonstration purposes. In practice, as mentioned above, many factors could impact the inference time, such as GPUs and efficient multi-processing procedures. The real inference time might differ, but the difference between the models holds, as we fairly compared all of the models in the same setting. On the basis of these results, we suggest using compressed or distilled BERT models [31] for real-time applications, especially when production servers do not have available GPUs.

The second primary finding is that the random forest model made more errors in sentence pairs of low similarity (similarity scores from 0 to 1), whereas BERT models made more errors on highly similar sentence pairs (similarity scores from 4 to 5). The random forest model cannot effectively capture the sentence semantics when a sentence pair shares consistent structures and similar words but distinct topics. In contrast, ClinicalBERT had an MSE of over 2.5 for highly similar sentence pairs, especially when different negation terms or the word order is switched. As mentioned above, the results also suggest that interannotator consistency may also impact MSE, showing the difficulty of relatedness-based tasks.

Limitations

The main limitation of this study is that the analysis was conducted using the ClinicalSTS data set alone. To the best of our knowledge, the data set is already the largest available sentence similarity data set in this domain. Other data sets, such

as BOSSES, are much smaller. We believe that it is critical to developing more sentence similarity data sets from other sources in the biomedical and clinical domains, which could expand our analysis and further improve the existing methods.

Another limitation is that the ClinicalSTS data set lacked user-level evaluations. The notion of relevance is context-dependent: sentence pairs with high similarity scores predicted by the models may not necessarily be considered relevant by users [32]. Previous studies demonstrated that the top sentences ranked by the top STS models were not the most relevant to users based on manual judgment [33]. Therefore, it is critical to conduct user-level assessments to understand whether STS models can facilitate information retrieval in practice, in addition to understanding the effectiveness and efficiency measures. We consider this as future work.

Comparison With Prior Work

Most existing studies focus on developing innovative methods to improve correlations in the testing set. Top-ranked methods are summarized in the overview papers on clinical STS challenge tasks [8,9], from traditional machine learning methods [30] to word and sentence embedding–based methods [20] and transformer-based methods [24]. Other studies further used advanced learning methods, such as representation fusion [34] and multitask learning [27]. The reported Pearson correlations range from 0.83 to 0.90, which is consistent with our study. Although it is exciting to further improve the state-of-the-art results, it is more critical to understand the effectiveness and efficiency of these models in depth, especially when the human-level correlation level is only moderate in these data sets.

Only 2 studies have compared the effectiveness of STS models in the biomedical and clinical domains [35,36]. Tawfik et al [35] compared the performance of a range of embeddings in sentence-based data sets (mostly classification-based applications, not STS) in the biomedical domain. Studies have shown that embeddings pretrained in biomedical and clinical corpora could achieve reasonable Pearson correlation scores, which is consistent with our study. However, these studies focused mainly on the Pearson correlations and did not consider model robustness or efficiency. Arguably, the latter is more critical to using STS models in practice.

Conclusions

In this postchallenge study, we comparatively analyzed the effectiveness and efficiency of 5 deep learning models in the ClinicalSTS data set. Although these models achieved high Pearson correlation scores, their robustness varied dramatically in terms of sentence pairs at different similarity levels, and BERT models have significantly longer inference times. In addition, the models achieved Pearson correlations of approximately 0.90 in this data set, whereas the human-level agreement was only moderate. Taken together, these observations make us cautious about the further improvement of this data set and argue for a more thorough evaluation of the model-generalization capability and user-level testing. We also call for community efforts to create more STS data sets from different perspectives to reflect the multifaceted notion of sentence relatedness, which will further improve the generalization performance of deep learning models.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine. This work was also supported by the National Library of Medicine of the National Institutes of Health under award 4R00LM013001. The authors thank Dr. Alexis Allot for helpful discussions on the sentence search pipeline in LitSense. They also thank the National Natural Language Processing Clinical Challenges/Open Health Natural Language Processing Consortium challenge task organizers for coordinating this shared task.

Conflicts of Interest

All authors are employees of the National Institutes of Health.

References

1. Prakoso DW, Abdi A, Amrit C. Short text similarity measurement methods: a review. Soft Comput 2021 Jan 03;25(6):4699-4723. [doi: 10.1007/s00500-020-05479-2]
2. Allot A, Chen Q, Kim S, Alvarez R, Comeau D, Wilbur W, et al. Litsense: Making sense of biomedical literature at sentence level. Nucleic Acids Res 2019 Jul 02;47(W1):594-599 [FREE Full text] [doi: 10.1093/nar/gkz289] [Medline: 31020319]
3. International Society for Biocuration. Biocuration: Distilling data into knowledge. PLoS Biol 2018 Apr 16;16(4):8 [FREE Full text] [doi: 10.1371/journal.pbio.2002846] [Medline: 29659566]
4. Chen Q, Leaman R, Allot A, Luo L, Wei C, Yan S, et al. Artificial intelligence in action: Addressing the covid-19 pandemic with natural language processing. Annu Rev Biomed Data Sci 2021 May 14;4(1):313-339. [doi: 10.1146/annurev-biodatasci-021821-061045] [Medline: 34465169]
5. Cer D. Semeval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). 2017 Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); August, 2017; Vancouver, Canada p. 1-14. [doi: 10.18653/v1/s17-2001]

6. Soganciogl G, Öztürk H, Özgür AB. Biosses: A semantic sentence similarity estimation system for the biomedical domain. Bioinform 2017 Jul 15;33(14):49-58 [FREE Full text] [doi: 10.1093/bioinformatics/btx238] [Medline: 28881973]
7. Wang Y, Afzal N, Fu S, Wang L, Shen F, Rastegar-Mojarad M, et al. Medsts: A resource for clinical semantic textual similarity. Lang Resour Eval 2018 Oct 24;54(1):57-72. [doi: 10.1007/s10579-018-9431-1]
8. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen P, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: Clinical semantic textual similarity. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. 2018 Presented at: BCB '18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 29 - September 1, 2018; Washington DC USA URL: https://www.researchgate.net/publication/327424883_Overview_of_BioCreativeOHNLP_Challenge_2018_Task_2_Clinical_Semantic_Textual_Similarity
9. Wang Y, Fu S, Shen F, Henry S, Uzuner O, Liu H. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. JMIR Med Inform 2020 Nov 27;8(11):11 [FREE Full text] [doi: 10.2196/23375] [Medline: 33245291]
10. Schober P, Boer C, Schwarte LA. Correlation coefficients. Anesth Analg 2018;126(5):1763-1768. [doi: 10.1213/ane.0000000000002864]
11. Mikolov T. Distributed representations of words and phrases and their compositionality. arXiv.org. 2013. URL: https://arxiv.org/abs/1310.4546 [accessed 2021-09-21]
12. Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).: Association for Computational Linguistics; 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October, 2014; Doha, Qatar. [doi: 10.3115/v1/d14-1162]
13. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec;5:135-146. [doi: 10.1162/tacl_a_00051]
14. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. Biowordvec, improving biomedical word embeddings with subword information and mesh. Sci Data 2019 May 10;6(1):52 [FREE Full text] [doi: 10.1038/s41597-019-0055-0] [Medline: 31076572]
15. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of word embeddings for the biomedical natural language processing. J Biomed Inform 2018 Nov;87:12-20 [FREE Full text] [doi: 10.1016/j.jbi.2018.09.008] [Medline: 30217670]
16. Chen Q, Lee K, Yan S, Kim S, Wei C, Lu Z. Bioconceptvec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. PLoS Comput Biol 2020 Apr 23;16(4):18 [FREE Full text] [doi: 10.1371/journal.pcbi.1007617] [Medline: 32324731]
17. 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; 2014; Bejing, China p. 1188-1196 URL: https://proceedings.mlr.press/v32/le14.html
18. Hill F, Cho K, Korhonen A. Learning distributed representations of sentences from unlabelled data. 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; June, 2016; San Diego, California p. 1367-1377. [doi: 10.18653/v1/n16-1162]
19. Pagliardini M, Gupta P, Jaggi M. Unsupervised learning of sentence embeddings using compositional n-gram features. arXiv.org. 2017 Mar 07. URL: https://arxiv.org/abs/1703.02507 [accessed 2021-09-21]
20. Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):73 [FREE Full text] [doi: 10.1186/s12911-020-1044-0] [Medline: 32349758]
21. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. Advances in Neural Information Processing Systems (NIPS). 2017. URL: https://www.aminer.org/pub/5c8b4d794895d9cbc67f0e33/attention-is-all-you-need [accessed 2021-09-21]
22. 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: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019); Jun 2 - 7, 2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
23. Radford A. Improving language understanding by generative pre-training. Open AI Codex. 2018 Jun 11. URL: https://openai.com/blog/language-unsupervised/ [accessed 2018-06-11]
24. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. Biobert: A pre-trained biomedical language representation model for biomedical text mining. Bioinform 2020 Feb 15;36(4):1234-1240 [FREE Full text] [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
25. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: 18th BioNLP Workshop and Shared Task; August, 2019; Florence, Italy p. 58-65. [doi: 10.18653/v1/w19-5006]

26. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv.org. 2019. URL: https://arxiv.org/abs/1904.03323 [accessed 2021-09-21]
27. Peng Y, Chen Q, Lu Z. An empirical study of multi-task learning on BERT for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 May 06 Presented at: 19th SIGBioMed Workshop on Biomedical Language Processing; July, 2020; Online p. 205-214. [doi: 10.18653/v1/2020.bionlp-1.22]
28. Shao Y. HCTI at SemEval-2017 Task 1: Use convolutional neural network to evaluate semantic textual similarity. In: Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).: Association for Computational Linguistics; 2017 Aug Presented at: 11th International Workshop on Semantic Evaluation (SemEval-2017); August, 2017; Vancouver, Canada p. 130-133. [doi: 10.18653/v1/s17-2016]
29. Chen Q, Peng Y, Lu Z. BioSentVec: Creating sentence embeddings for biomedical texts. In: Proceedings of the IEEE International Conference on Healthcare Informatics (ICHI). 2019 Presented at: The Seventh IEEE International Conference on Healthcare Informatics (ICHI 2019); June 10-13, 2019; Xi'an, China. [doi: 10.1109/ichi.2019.8904728]
30. Chen Q, Du J, Kim S, Wilbur WJ, Lu Z. Combining rich features and deep learning for finding similar sentences in electronic medical records. In: Proceedings of the BioCreative/OHNLP Challenge. 2018 Presented at: BioCreative/OHNLP Challenge; August 29 - September 1, 2018; Washington DC, USA URL: https://www.researchgate.net/publication/327402060_Combining_rich_features_and_deep_learning_for_finding_similar_sentences_in_electronic_medical_records
31. Tang R, Lu Y, Liu L, Mou L, Vechtomova O, Lin J. Distilling task-specific knowledge from bert into simple neural networks. arXiv.org. 2019. URL: https://arxiv.org/abs/1903.12136 [accessed 2021-09-21]
32. Saracevic T. The notion of relevance in information science: everybody knows what relevance is. But, what is it really? In: Synthesis Lectures on Information Concepts, Retrieval, and Services. Williston, United States: Morgan & Claypool; Sep 06, 2016:i-109.
33. Chen Q, Kim S, Wilbur W, Lu Z. Sentence similarity measures revisited: Ranking sentences in pubmed documents. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics.: Association for Computing Machinery; 2018 Presented at: BCB '18: 9th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics; August 29 - September 1, 2018; Washington DC, United States p. 531-532. [doi: 10.1145/3233547.3233640]
34. Xiong Y, Chen S, Qin H, Cao H, Shen Y, Wang X, et al. Distributed representation and one-hot representation fusion with gated network for clinical semantic textual similarity. BMC Med Inform Decis Mak 2020 Apr 30;20(Suppl 1):72 [FREE Full text] [doi: 10.1186/s12911-020-1045-z] [Medline: 32349764]
35. Tawfik NS, Spruit MR. Evaluating sentence representations for biomedical text: methods and experimental results. J Biomed Inform 2020 Apr;104:103396 [FREE Full text] [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
36. Antunes R, Silva JF, Matos S. Evaluating semantic textual similarity in clinical sentences using deep learning and sentence embeddings. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. 2020 Presented at: SAC '20: Proceedings of the 35th Annual ACM Symposium on Applied Computing; March 30 - April 3, 2020; Brno Czech Republic p. 662-669. [doi: 10.1145/3341105.3373987]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
CNN: Convolutional Neural Network
GPU: graphics processing unit
n2c2: National Natural Language Processing Clinical Challenges
OHNLP: Open Health Natural Language Processing
STS: semantic textual similarity

Edited by Y Wang; submitted 22.01.21; peer-reviewed by BJ Webb-Robertson, M Manzanares; comments to author 16.03.21; revised version received 06.08.21; accepted 06.08.21; published 30.12.21.

Please cite as:
Chen Q, Rankine A, Peng Y, Aghaarabi E, Lu Z
Benchmarking Effectiveness and Efficiency of Deep Learning Models for Semantic Textual Similarity in the Clinical Domain: Validation Study
JMIR Med Inform 2021;9(12):e27386
URL: https://medinform.jmir.org/2021/12/e27386
doi:10.2196/27386
PMID:34967748

Original Paper

Differential Biases and Variabilities of Deep Learning–Based Artificial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study

Dongchul Cha[1], MD; Chongwon Pae[2,3], PhD; Se A Lee[1], MD; Gina Na[1], MD; Young Kyun Hur[1], MD; Ho Young Lee[1], MD; A Ra Cho[1], MD; Young Joon Cho[4], MD; Sang Gil Han[4], MD; Sung Huhn Kim[1], MD, PhD; Jae Young Choi[1*], MD, PhD; Hae-Jeong Park[2,3*], PhD

[1]Department of Otorhinolaryngology, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Center for Systems and Translational Brain Sciences, Institute of Human Complexity and Systems Science, Yonsei University College of Medicine, Seoul, Republic of Korea

[3]Graduate School of Medical Science, Brain Korea 21 Project, Department of Nuclear Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[4]Department of Emergency Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

[*]these authors contributed equally

Corresponding Author:
Hae-Jeong Park, PhD
Center for Systems and Translational Brain Sciences
Institute of Human Complexity and Systems Science
Yonsei University College of Medicine
50-1 Yonsei-ro
Seoul, 03722
Republic of Korea
Phone: 82 2 2228 2363
Email: parkhj@yuhs.ac

Abstract

Background: Deep learning (DL)–based artificial intelligence may have different diagnostic characteristics than human experts in medical diagnosis. As a data-driven knowledge system, heterogeneous population incidence in the clinical world is considered to cause more bias to DL than clinicians. Conversely, by experiencing limited numbers of cases, human experts may exhibit large interindividual variability. Thus, understanding how the 2 groups classify given data differently is an essential step for the cooperative usage of DL in clinical application.

Objective: This study aimed to evaluate and compare the differential effects of clinical experience in otoendoscopic image diagnosis in both computers and physicians exemplified by the class imbalance problem and guide clinicians when utilizing decision support systems.

Methods: We used digital otoendoscopic images of patients who visited the outpatient clinic in the Department of Otorhinolaryngology at Severance Hospital, Seoul, South Korea, from January 2013 to June 2019, for a total of 22,707 otoendoscopic images. We excluded similar images, and 7500 otoendoscopic images were selected for labeling. We built a DL-based image classification model to classify the given image into 6 disease categories. Two test sets of 300 images were populated: balanced and imbalanced test sets. We included 14 clinicians (otolaryngologists and nonotolaryngology specialists including general practitioners) and 13 DL-based models. We used accuracy (overall and per-class) and kappa statistics to compare the results of individual physicians and the ML models.

Results: Our ML models had consistently high accuracies (balanced test set: mean 77.14%, SD 1.83%; imbalanced test set: mean 82.03%, SD 3.06%), equivalent to those of otolaryngologists (balanced: mean 71.17%, SD 3.37%; imbalanced: mean 72.84%, SD 6.41%) and far better than those of nonotolaryngologists (balanced: mean 45.63%, SD 7.89%; imbalanced: mean 44.08%, SD 15.83%). However, ML models suffered from class imbalance problems (balanced test set: mean 77.14%, SD 1.83%; imbalanced test set: mean 82.03%, SD 3.06%). This was mitigated by data augmentation, particularly for low incidence classes, but rare disease classes still had low per-class accuracies. Human physicians, despite being less affected by prevalence, showed high interphysician variability (ML models: kappa=0.83, SD 0.02; otolaryngologists: kappa=0.60, SD 0.07).

Conclusions: Even though ML models deliver excellent performance in classifying ear disease, physicians and ML models have their own strengths. ML models have consistent and high accuracy while considering only the given image and show bias toward prevalence, whereas human physicians have varying performance but do not show bias toward prevalence and may also consider extra information that is not images. To deliver the best patient care in the shortage of otolaryngologists, our ML model can serve a cooperative role for clinicians with diverse expertise, as long as it is kept in mind that models consider only images and could be biased toward prevalent diseases even after data augmentation.

(JMIR Med Inform 2021;9(12):e33049) doi:10.2196/33049

KEYWORDS

human-machine cooperation; convolutional neural network; deep learning, class imbalance problem; otoscopy; eardrum; artificial intelligence; otology; computer-aided diagnosis

Introduction

Machine learning (ML) based on deep learning (DL) in medical imaging is developing at a rapid pace, to fill the gap between the capacity of specialists interpreting the images and the need for interpreted images. Many studies [1-6] show the possibility that the performance of image classification is on par or better than that of medical specialists in terms of accuracy. Despite the promising results of these studies, characteristics of DL have not been thoroughly evaluated and compared with human experts, particularly in the domain of clinical practice. In tasks such as medical image diagnosis, where accountability is an important issue, cooperation between human experts and ML models is necessary [1]. To foster cooperation between humans and machines, the characteristics of human intelligence (HI) and DL-based artificial intelligence (AI) should be specified at the individual and systemic levels.

The class imbalance in real-world clinics is a big challenge in data-driven ML. Different numbers of samples in various classes due to imbalanced incidences inherent in the human population are expected to induce biases toward high incident classes during the training process.

Conversely, human medical experts learn in-depth by experiencing limited numbers of cases, thus have less bias for classes of different sizes [7]. However, clinical experience differs among clinicians, and every clinician has their own classification biases, that is, strengths and weaknesses in classifying certain diseases [8]. Due to the bias induced by individual experience, physicians may have large interindividual variability. Meanwhile, ML models are statistically biased based on the amount of data but show consistent performance among different models [9]. Despite general speculations, these 2 biases for data size for each class and interindividual variation due to differential (small sample–biased) experiences have not been directly evaluated in the clinical diagnostic setting.

In this study, we investigated the differential characteristics of ML models and human experts concerning class imbalance bias and interrater variability. For this, we use as the example the classification of ear and mastoid disease using otoendoscopic images. Ear and mastoid diseases are common in, but not limited to, developing countries in Southeast Asia, Western Pacific regions, and Africa [10]. However, otolaryngologists are shorthanded in many developing countries, with as few as <1 otolaryngologist per a million people in 64% of African counties [11]. Therefore, nonotolaryngologists in primary care are likely to see patients with these diseases in clinics, and they must play a role in managing ear diseases, particularly in areas with limited access to otolaryngologists. However, nonotolaryngologists are prone to misdiagnosing otitis media, which is a major part of ear disease [11-13]. Evaluating ear disease involves careful history taking and physical examination using conventional otoscopy or otoendoscopy. The initial impression of otoscopy is an essential gateway to diagnosis and treatment.

One of the domain-specific challenges in ear disease classification, as in other medical fields, is the class imbalance problem discussed earlier. This problem may affect both clinicians and ML models but possibly more so ML models. Because ear diagnosis is conducted by clinicians with diverse levels of expertise, the variability of individual performance is apparent in this field [14,15].

To investigate and compare the effect of the class imbalance problem between human physicians and ML models as well as interindividual variability, we evaluated the diagnostic rate and interrater reliability of otoendoscopic images among 3 groups: otolaryngologists (2 specialists and 4 residents), nonotolaryngologists (2 family medicine specialists, 2 emergency medicine specialists, and 5 general practitioners), and 13 convolutional neural network (CNN)–based classification models in both balanced and imbalanced test sets, each containing 300 otoendoscopic images. We also examined the dependency of the accuracy on the prevalence of each class in the machines compared with that of human experts. The class imbalance problem was evaluated concerning diverse data augmentation strategies generalizable for most CNN-based classification models to overcome the aforementioned class imbalance problem. We also evaluated the effect of the augmentation strategy in improving classification accuracy according to the incidence of the disease. All these evaluations were conducted by optimizing our previous automated diagnosis system [9]. Furthermore, we sought the possibility of using our classification system as a virtual otolaryngologist to assist physicians by comparing the accuracy and likelihood of diagnosis between our classification system and otolaryngologists.

Methods

Patient Data Selection and Acquisition

Digital otoendoscopic images from patients who visited the outpatient clinic in the Department of Otorhinolaryngology at Severance Hospital, Seoul, South Korea from January 2013 to June 2019 were used. A total of 22,707 otoendoscopic images routinely taken using different otoendoscopic cameras by otolaryngology residents, faculty, or experienced nurses were reviewed for labeling. The image resolution was 640 x 480 pixels in the DICOM format. We excluded postsurgerical status photos, duplicate images, images that were significantly out of focus or fuzzy, and otoendoscopic images from the same patient's follow-up data without changes in the diagnosis. We aggressively excluded similar images if an image was taken multiple times at slightly different angles; we selected only one of the images. As a result, 7500 otoendoscopic images were selected for labeling. This study was approved by the Severance Hospital Institutional Review Boards (IRB No 2019-0467-001). Written informed consent was obtained from physician participants. All methods complyed with the Declaration of Helsinki.

Analysis and Labeling of Otoendoscopic Images

Otoendoscopic photos containing eardrums and the external auditory canal (EAC) were classified into 6 categories to cover all diseases based on the *Color Atlas of Endo-Otoscopy* [16]: (1) normal eardrum and EAC including healed perforation and tympanosclerosis; (2) tumorous condition, in which there are tumors in the middle ear, EAC, or cerumen impaction; (3) otitis media with effusion; (4) myringitis or otitis externa; (5) perforated eardrums; (6) attic retraction or middle ear atelectasis. Internally, there were more subclasses, but we consequently merged those subclasses into the 6 aforementioned classes because we could not acquire an adequate number of sample sizes of smaller subclasses. Since the goal of the diagnosis system is to offer an appropriate treatment strategy in real-world clinics, the label was constructed considering both required treatment and the similarity of physical findings.

Often, there could be multiple etiologies present in 1 otoendoscopic image. For example, attic retraction with middle ear effusion could be present. In such cases, the image was labeled as attic retraction according to our labeling priority. This priority was determined by the certainty of disease and possible need for surgery.

To ensure the ground-truth label was correct, we applied additional steps in labeling, since the accuracy of otoscopy by a single physician may only be 75% [17]. First, all images were double-checked by reviewing the patient's diagnosis in the electronic medical record by the attending physician at the time, who had at least 10 years of clinical experience in a tertiary referral center. Second, if the otoendoscopic image was not trivial, even after reviewing the medical records, additional test results (audiological tests including pure-tone audiometry and impedance audiometry, radiological tests including computed tomography, magnetic resonance imaging) were considered for labeling the ground truth. Last, if the first author could not agree or make an appropriate impression on the otoendoscopic image even after combining medical records and additional tests, the picture was discarded. An in-house graphic user interface software built with MATLAB2019a (MathWorks Inc, Natick, MA) was used for manual labeling.

Supervised Training of CNN Models for EAC Data With Transfer Learning

Public CNN models were pretrained with the ImageNet database [18] to classify 1000 natural objects that served as a base model for transfer learning of otoendoscopic images. Pareto-efficient models were chosen to be transferred to this study domain. They were ResNets [19] (ResNet101, ResNet152), InceptionV3 [20], InceptionV4 [21], Inception-ResNet-V2 [21], VGG-19 with batch normalization [22], SENet [23], DenseNet [24], and NASNet [25,26]. Those models were modified to classify 6 categories of otoendoscopic images by replacing the last fully connected layer of each model with a layer of 6 fully connected output nodes. For model optimization, Adaptive Moment Estimation (ADAM) [27] with a batch size of 32 was used. Larger batch sizes were not used according to a study reporting the advantage of smaller batch sizes [28]. We trained for a total of 20 epochs with differential learning rates. The initial learning rate was 0.01 in the last transferred layer for 5 epochs. After 5 epochs, fine-tuning was done: All the layers were trained for 7 epochs with a discriminative learning rate, ranging from $1x10^{-4}$ in the last layer to $1x10^{-6}$ in the first layer. Afterward, we trained for 7 epochs with a learning rate of $1x10^{-9}$ in the last layer and $3.3x10^{-10}$ in other layers. To prevent overfitting, affine transformations of images were applied. A horizontal flip, rotation of up to 20 degrees, random scales between 0.8 and 1.2, change of lighting up to 20%, and a random symmetric warp of magnitude between –0.2 and 0.2 were randomly applied with a probability of 75% on every epoch. Model construction, training, validation, and testing were implemented using Pytorch [29] with the Fastai library [30].

Comparison of the Accuracy of the Models With Diverse Training Settings

Comparison of Model Construction and Performance According to Training Sample Size

Among a total of 7500 otoendoscopic images, 7200 images (300 mutually exclusive images were left out for testing in both balanced and imbalanced scenarios; Table 1) were used for training in 20 epochs. To maximize available data for training, we included data from other test sets into the training set; that is, we put the imbalanced test dataset into the training set when evaluating in the balanced testing environment and vice versa when evaluating in the imbalanced testing environment.

Table 1. Composition of the training and test sets as well as labels, sorted by labeling priority.

Classification	Number of images		
	Training (n=6900), n (%)	Test-balanced[a] (n=300), n (%)	Test-imbalanced[b] (n=300), n (%)
(1) Tympanic perforation	1793 (26.99)	50 (16.77)	51 (17.00)
(2) Attic retraction/atelectasis	521 (7.56)	50 (16.77)	20 (6.67)
(3) Myringitis/otitis externa	256 (3.71)	50 (16.77)	15 (5.00)
(4) Otitis media with effusion	506 (7.33)	50 (16.77)	29 (9.67)
(5) Tumors	285 (4.13)	50 (16.77)	18 (6.00)
(6) Normal	3539 (51.29)	50 (16.77)	167 (55.67)

[a]All classes are distributed equally.

[b]Classes are distributed proportionally to the training set.

We chose random image samples using different random seeds 5 times to flatten accuracy fluctuations. Performance according to training sample size was evaluated to verify the significance of the larger training sample size: 10% (720 images), 25% (1800 images), 50% (3600 images), 90% (6480 images), and 100% (7200 images).

Strategies to Overcome Class Imbalance Between Labels

Class imbalance was inevitable due to the diverse incidence of various ear diseases. To mitigate this problem, 3 strategies were incorporated in training: oversampling, the mixup [31] method, and focal loss [32] as the loss function (focal loss with $\gamma = 1$). Oversampling was done by copying images in the smaller classes to a level equivalent to the largest class, combined with affine transformations of images. Images of diseases other than normal eardrums were oversampled to reach the number of normal eardrum images in the current database. Images of otitis media with effusion and attic retractions were augmented approximately 6-fold. The images of myringitis and tumors required almost 10-fold oversampling. Mixup and focal loss are described in detail in Multimedia Appendix 1.

We tested 12 models with 8 different configurations (baseline, with and without oversampling, focal loss, and mixup) resulting in a total of 12 x 2 x 2 x 2 = 96 CNN-based ML model variants.

Evaluation of the ML Model Accuracy and Similarities in Prediction Tendency in Both Balanced and Imbalanced Test Sets

After fine-tuning various CNN-based ML models, the accuracies of all models were evaluated in both balanced and imbalanced testing scenarios (Table 1). The first, balanced, 300-image set consisted of 50 images for each label, which is different from the incidence ratio in clinical settings but better suited for measuring accuracies. The second, imbalanced, 300-image set contained different numbers of images with each label based on its prevalence in the database, which may represent the proportion of disease in real-world clinics, particularly a tertiary referral hospital. Also, the likelihood of diagnosis between different ML models was evaluated using the Fleiss kappa method [33]. The kappa (κ) scores were interpreted as follows: $\kappa<0$ as poor, 0.01-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1 as almost perfect agreement [34].

Evaluation of Human Physicians' Diagnoses: Accuracy and Variability

A computerized online questionnaire consisting of 2 sets of 300 questions, identical to the ML model's balanced and imbalanced test sets (600 images in total, Table 1), was presented to 14 participants in 3 groups: 6 otolaryngologists (2 otolaryngologists, 4 otolaryngology residents), 8 nonotolaryngologists who had previous exposure to otoscopy (2 emergency medicine specialists, 2 family medicine specialists), and 4 general practitioners). Informed written consent was obtained from all participants.

All participants answered the questionnaire in the same order. Participants were requested to answer according to the same labeling priority as in ML models if more than one pathology was present in the given image. Along with the diagnosis, the participants were asked to rate the confidence of their diagnosis on a scale of 1 (not confident) to 10 (very confident). The participants were not told whether the set was balanced or imbalanced, since it might have provided additional clinical suspicion of less common disease entities.

Interrater agreement among individual groups was calculated using the aforementioned Fleiss kappa method [33]. Spearman correlation analyses were also performed to check the possible relationships between confidence and accuracy of diagnosis to determine whether higher confidence is associated with better accuracy.

Comparison of Diagnostic Performance and Tendency Between Physicians and ML Models

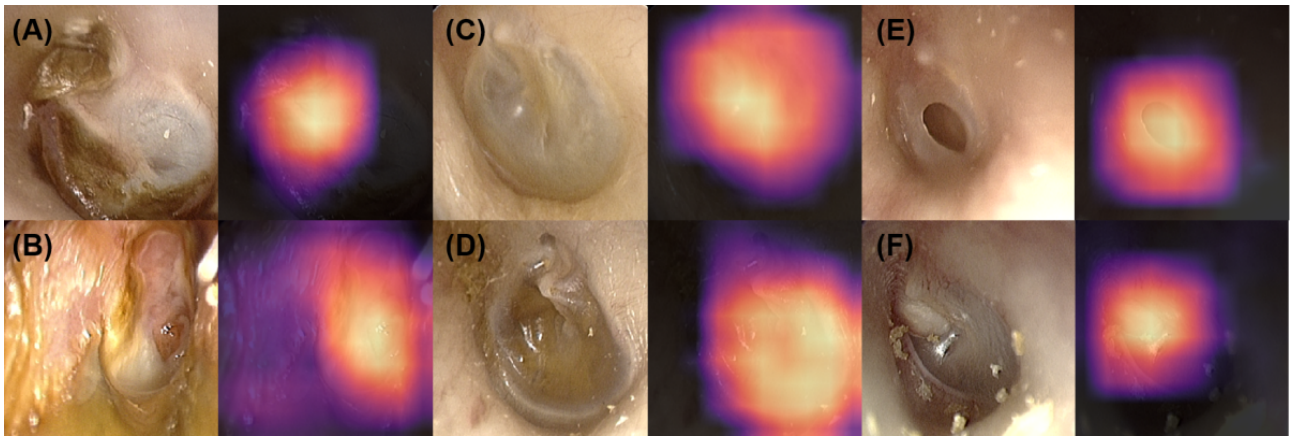
All the answers, which were provided in identical order, from the human physicians and ML models were lined up to compare the accuracy. We evaluated the differences in the classification pattern depending on the class prevalence between physicians and ML models in both balanced and imbalanced test sets. We measured the likelihood of the ML model's diagnosis to that of human physicians by comparing kappa values. We also compared the per-class accuracy, precision, recall, and F1 scores between physicians and ML models. We then analyzed the differential effects of class prevalence in accuracy and prediction counts using linear regression analysis.

Results

Training and Test Sets

We used a total of 6900 otoendoscopic images from 6 classes for training (Table 1). The training dataset was imbalanced, reflecting the prevalence of ear disease. Although the dataset was obtained based on a tertiary referral center, therefore having rich pathologic cases, normal classes were substantially common. The testing environment consisted of 2 different settings: (1) balanced test set (300 sample images), consisting of 6 classes with 50 images each, without considering the prevalence of ear diseases and (2) imbalanced test set (300 samples), each class distributed proportionally to the training dataset. Figure 1 displays representative classes and their activation heatmaps. The classification system could focus on important areas of eardrums and EACs. For attic retraction, the DL model focused on pathologic attic areas of the eardrum. When EACs were wet due to inflammation of the middle or external ear, it was visible in the heatmap. Normal and middle ear effusions have the same area of interest, mainly the eardrum and the middle ear cavity, which was correctly depicted by the classification system. Perforation of the tympanic membrane was visualized by the heatmap, as well as middle ear tumors inside the tympanic membrane (Figure 1).

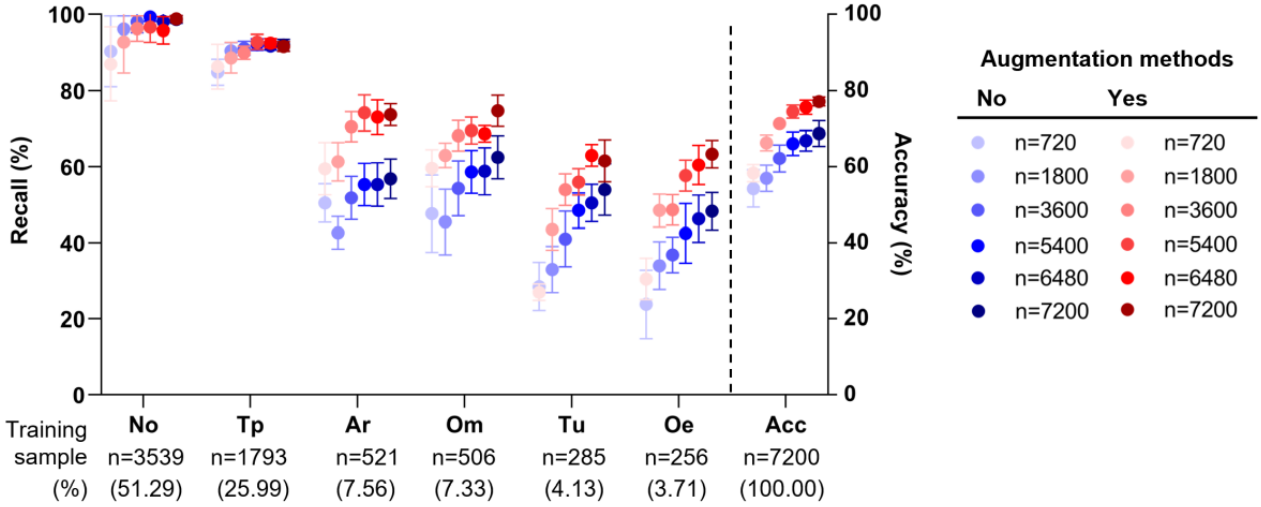
Figure 1. Representative class and their activation heatmap (Grad-CAM): (A) attic retraction), (B) myringitis or otitis externa, (C) normal findings, (D) otitis media with effusion, (E) tympanic perforation, (F) middle ear or external ear canal tumors.



ML Model Performance Over Different Numbers of Training Samples, the Class Imbalance Problem, and Modifications

When testing with the baseline model (without adjustment of class imbalance in training), the overall average accuracy was 82.78% in the imbalanced (according to disease prevalence) test set. However, in the balanced test set, the overall accuracy was 68.69% (chance level: 16.7%), substantially inferior to the accuracy of 82.78% for the imbalanced testing data. To mitigate the class imbalance problem, we retrained a classification model using oversampling, mixup, and focal loss. We tested every combination of these strategies under the balanced testing environment. Applying all 3 strategies in the training phase had a synergistic effect, achieving an average of 8.41% gain (average accuracy: 77.14% vs 68.69%) in the balanced test set while compromising 0.75% in the imbalanced test set. Especially, oversampling was universally beneficial (Multimedia Appendix 2). The augmented classifier gained more per-class accuracy for classes with fewer samples, such as attic retractions, than the baseline model, leading to better overall accuracy in the balanced test set (n=7200; Figure 2; additional example results of both test sets with a Resnet101-based classifier available in Multimedia Appendix 3).

Figure 2. Per-class recall and overall classification accuracy (bars = 95% CI) for classes according to the number of training samples and augmentation, trained with 12 different convolutional neural network models and tested on the balanced test set. Acc: overall accuracy; Ar: attic retraction, destruction; No: normal; Oe: myringitis or acute otitis externa; Om: otitis media with effusion; Tp: tympanic perforation; Tu: middle or external ear canal tumors or cerumen impaction.

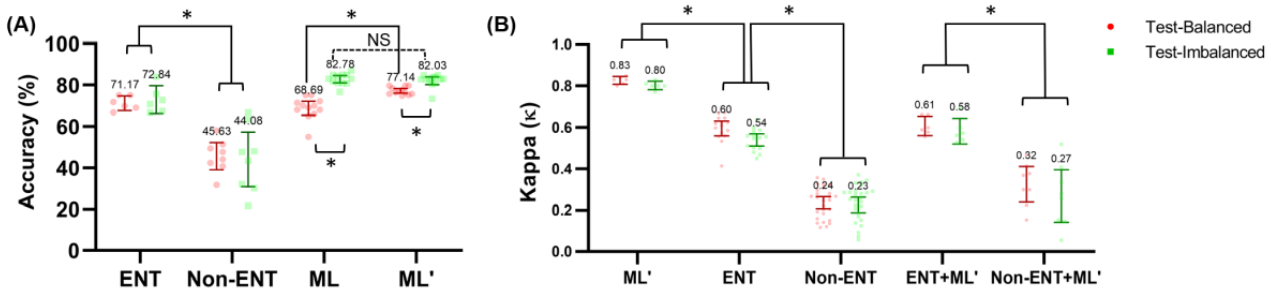


To explore the relationship between the classification bias and the size of the training dataset in detail, we compared the classification performance over different numbers of training samples when tested in a balanced testing environment. The overall accuracy increased with a higher number of samples. The adjustment for the class imbalance during the training steps improved the performance, particularly for classes with fewer training samples (Figure 2). For classes with a high incidence rate, there were no significant gains by augmentation as they already reached a plateau of accuracy, not to mention the oversampling method synthesizes more images for smaller classes to match the most common, "normal," class. Nevertheless, augmenting images (affine transformations) for rare classes did not yet reach a saturated accuracy as the number of total training samples increased.

AI Versus HI in Per-Class Accuracy and Interrater Variability

The diagnostic accuracy of the 2 test sets was evaluated separately (Table S2 in Multimedia Appendix 4; additional metrics including precision, recall, and F1 scores are in Multimedia Appendix 4). All participants, including prediction models, assessed the same collection of images in the same order to rule out bias caused by different questionnaire layouts. Otolaryngologists (n=6) significantly outperformed nonotolaryngologists (n=8) in both balanced (mean 71.17%, SD 3.37% vs mean 45.63%, SD 7.90%; Mann-Whitney U=0; P<.001) and imbalanced (mean 72.84%, SD 6.41% vs mean 44.08%, SD 15.84%; Mann-Whitney U=0.5; P=.001) test sets. Our fine-tuned CNN-based ML models (n=12) tended to be better than otolaryngologists (n=6) in both imbalanced (mean 82.03%, SD 3.06% vs mean 72.84, SD 6.41%; Mann-Whitney U=10.50; P=.014) and balanced (mean 77.14%, SD 1.84% vs mean 71.17%, SD 3.37%; Mann-Whitney U=3; P<.001) test sets and outperformed nonotolaryngologists in both test sets (Figure 3A).

Figure 3. Mean (A) overall diagnostic accuracy and (B) Fleiss generalized kappa for interrater reliability (error bars = 95% CI); the predictions by the ResNet152-based deep learning model were assumed to be a human rater. ENT: otolaryngologists; ENT+ML': machine learning model plus otolaryngologists; ML: baseline machine learning models; ML': augmented machine learning models; Non-ENT: nonotolaryngologists; Non-ENT+ML': machine learning model plus nonotolaryngologists; NS: not statistically significant. *P<.001 (Mann-Whitney test: ENT vs Non-ENT; Wilcoxon matched-pairs signed-rank test: ML vs ML').



Compared with nonotolaryngologists, ML models had better accuracy in all classes. Compared with otolaryngologists, ML models were better at predicting normal ears, tympanic perforations, and attic retractions, which were more prevalent in the training dataset. The diagnosis rate of otitis media with effusion and myringitis was similar between prediction models and otolaryngologists. For classifying tumorous conditions, otolaryngologists were better than prediction models in the

balanced test set (Table S2 in Multimedia Appendix 4). The overall accuracy for all physicians was not significantly different between the balanced and imbalanced test sets, while both augmented (n=12; median 5.3; *P*=.001; Wilcoxon matched-pairs signed-rank test) and baseline (n=12; median 13.3; *P*<.001; Wilcoxon matched-pairs signed-rank test) ML models had significantly higher accuracy in the imbalanced test set (Figure 3A). Of note, augmented ML models had gained significant accuracy in the balanced test set (n=12; median 8.3; *P*<.001; Wilcoxon matched-pairs signed-rank test) without loss of accuracy in the imbalanced test set (n=12; median 0.8; *P*=.28; Wilcoxon matched-pairs signed-rank test) compared with ML models without augmentation.

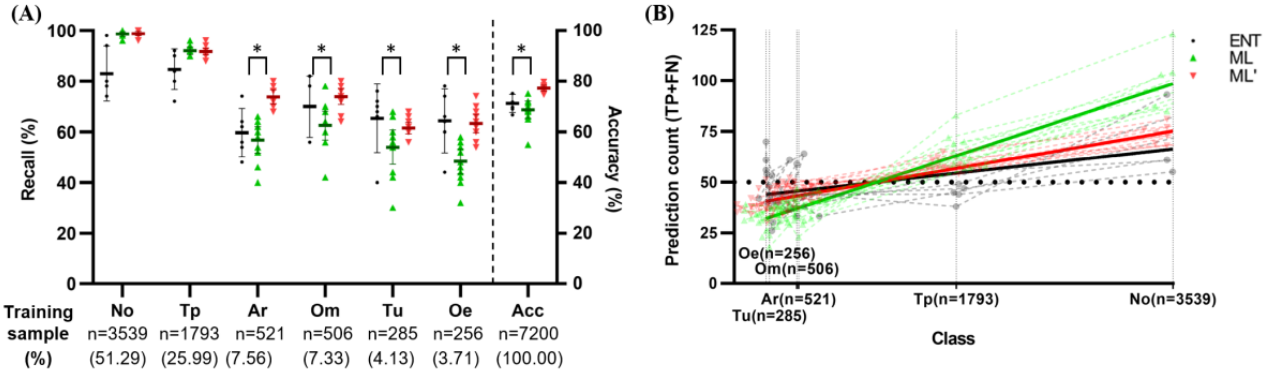
Regarding variance in accuracy, ML models had similar prediction results across different models, resulting in a low SD (1.76%), which was much lower than that of the otolaryngology specialists (5.86%) and nonotolaryngologists (14.82%). The results of the Fleiss generalized kappa as a measure of interrater reliability are presented in Figure 3B. Between ML models, κ scores ranged between 0.77 and 0.85, indicating a substantial diagnostic similarity among ML models. The κ score was >0.60 between 2 otolaryngology specialists and mostly >0.50 between all otolaryngology specialists and residents, which corresponds to moderate agreement between them. However, it was mostly <0.30 between nonotolaryngologists, which may be interpreted as fair agreement between these physicians. The predictions by the ML models were more likely to resemble those of otolaryngologists than nonotolaryngologists, showing similarity to otolaryngologists (Figure 3B; κ=0.5947, SD 0.05, n=12 vs κ=0.2966, SD 0.13, n=16; *P*<.001; Mann-Whitney U test).

Using the 4 top-performing models (ResNet152, DPN92, InceptionV4, and Densenet201), we constructed an ensemble classifier by adding and taking the maximum arguments following the softmax activation function in each classifier. Using this approach, we were able to gain an average of 1.83% in the balanced dataset and 3.5% in the imbalanced dataset, reaching 80.33% and 86.67% overall accuracy, respectively (Table S2 in Multimedia Appendix 4). The ensemble classifier of the different models outperformed any other CNN-based classifier alone in overall accuracy and proved to be a stable model for final prediction. Indeed, ensembling had a positive but, at the same time, limited effect in enhancing the overall accuracy because of diagnostic similarity, as indicated by high κ scores between models.

AI Compared With HI in Class Prevalence and Size of the Training Dataset

In otolaryngologists, accuracies tended to be stable regardless of sample sizes, whereas ML models showed a bias towards prevalent classes (Figure 4). Also, the augmentation method showed significantly improved accuracies in minor classes (attic retraction: n=12, median 15.0, *P*<.001; otitis media with effusion: n=12, median 13.0, *P*=.005; middle or external ear canal tumors or cerumen impaction: n=12, median 6.0, *P*=0.01; myringitis or acute otitis externa: n=12, median 13.0, *P*<.001; Wilcoxon matched-pairs signed rank tests). Otolaryngologists had a higher variance in the accuracies compared with the augmented ML models in prevalent classes (normal, tympanic perforation) and overall accuracy. We additionally analyzed the count of predicted samples, which corresponds to true-positive and false-negative predictions, for each class of the balanced test set. Each classification had 50 occurrences in the set, so ideally, the count of predicted samples (true positives and false negatives) should be 50, which is drawn as the dotted line in Figure 4B.

Figure 4. In the balanced test set, (A) per-class recall and overall accuracy (bars indicate 95% CI) and (B) prediction counts in individual classes (the dotted line at 50 indicates the sample size of the balanced test set for each class; x axis is on a logarithmic scale). Classes are listed left to right by descending number of training samples. Each class had 50 samples in the balanced test set (a total of 300 samples for all 6 classes). Nonotolaryngologists had too high variations and low accuracies and were not plotted. ENT: Y intercept=42.14 (95% CI 39.14-45.24), slope=0.006836 (95% CI 0.004805-0.008939), pseudo R-squared=0.3262; ML': Y intercept=37.89 (95% CI 35.77-40.07), slope=0.01053 (95% CI 0.008981-0.01211), pseudo R-squared=0.8665; ML: Y intercept=26.68 (95% CI 24.73-28.69), slope=0.02028 (95% CI 0. 01861-0.02198), pseudo R-squared=0.9167. Acc: overall accuracy; Ar: attic retraction; ENT: otolaryngologist; FN: false negative; ML: baseline machine learning models; ML': augmented machine learning models; No: normal; Oe: myringitis or acute otitis externa; Om: otitis media with effusion; Tp: tympanic perforation; TP: true positive; Tu: middle or external ear canal tumors or cerumen impaction. *P<.01 (Wilcoxon matched-pairs signed rank test).



ML models showed more bias towards the number of training data, as more prevalent classes tended to have more than 50 counts (above the dotted line; normal class was above the line and therefore overly diagnosed), while rarer classes such as myringitis or acute otitis externa had a lower count (below the dotted line; underdiagnosed). Different classification tendencies of humans and machines were evaluated with respect to their dependency on class prevalence. The Poisson regression analysis for the correlation between each class's number with the corresponding number of predictions showed a significantly

different slope between otolaryngologists, augmented ML models, and baseline ML models (Figure 4B; slope: 0.021 for ML, 0.011 for ML', and 0.007 for otolaryngologists). The likelihood ratio test with the null hypothesis had one curve for all data sets, and the alternative hypothesis had a different curve for each data set. The likelihood ratio between baseline ML models and the augmented ML models was 76.36 (P<.001), and the likelihood ratio between the augmented ML models and humans was 7.958 (P=.019). Differing slopes indicated that the ML models tended to produce more likely predictions based on the number of training samples.

Of note, otolaryngologists' predictions were not well fitted linearly because of individual differences in prediction (pseudo R-squared=0.3262). While the augmented ML model had mitigated the class imbalance problems, it still preferred prevalent classes, which was not apparent with otolaryngologists.

Discussion

Principal Findings

The main implications of this study are 3-fold: (1) Work by HI and AI shows different behaviors (prevalence dependency and interrater variability); (2) data augmentation reduces the class imbalance problem but the result is different according to the sample sizes of each class, requiring a certain amount of data samples for the rare class to achieve a reliable level; and (3) considering the high accuracy comparable to otologists and high variations in diagnostic performance by site clinicians, our ML model may act as a virtual otoendoscopic image analysis consultant, as long as clinicians consider that this ML model considers only images and there are potential biases in the ML models toward prevalence.

First, we showed that machines work in different ways than human knowledge, which is exemplarily reflected in the effects of class imbalance. As expected, ML models showed bias toward higher prevalent samples in the training set, but lower interrater (or ML model) variations. In contrast, human experts showed high interrater variations in their classifications but no prevalence-dependent biases. For example, the normal class is diagnosed when all other pathologies are excluded; hence, it is inherently difficult to diagnose despite its extensive prevalence. Meanwhile, cerumen impaction and tympanic perforation were less prevalent in the dataset, but they were classified correctly more times than the normal class by the human raters because of the obvious findings. Attic retractions and otitis media with effusions were subtle in many cases; hence, they were diagnosed with lower accuracy (Figure 4A). Therefore, for physicians, the difficulty lies mainly in class-specific abstract rules, which the data-driven ML model does not detect.

Second, although the class imbalance problem was mitigated by combining strategies in the training phase (oversampling, mixup, and focal loss), it had less effect for prevalent diseases but more for rare diseases (Figure 2). For the data-driven approach using ML, finding the hyperspace of features that covers within-class diversity, different from the other classes, is not trivial. ML attempts to find within-class diversity using imaging features based on statistics, which demands a large sample size to capture within-class variability. Indeed, in ML models, a higher number of samples in training produced better accuracies and reduced model variability (Figure 2), which is in line with the results of a previous study [9]. In reality, due to low incidence, we lacked a sufficient number of data samples for less prevalent diseases. Data augmentation improves the overall accuracy and recall of individual classes, especially for less prevalent classes. However, data augmentation was performed by manipulating the given dataset, which limited its diversity within images for rare classes compared with that of prevalent classes. Therefore, having more actual data for training is still essential for higher performance, particularly for rare classes. Often, datasets contain an abundance of normal and common disease classes and lack uncommon diseases. It is a general problem in the field of medical imaging, especially when diseases are rare and obtaining sufficient samples is difficult [35].

Third, our ML model showed the possibility of acting as a physician's assistant in real-world clinics. Inconsistent performance in humans was apparent, especially in the group of nonotolaryngologists (κ=0.24, 95% CI 0.21-0.26) compared with ML models (κ=0.83, 95% CI 0.81-0.84). Physicians often overestimated their skills despite the variance in their diagnostic capabilities, leading to faulty and inconsistent clinical information delivered to patients. Meanwhile, machines sometimes produced errors in trivial cases, even if their overall accuracies were expected to be on par or better than those of otolaryngologists. When making diagnostic suggestions, physicians' decisions should be taken into account to compensate for faulty ML suggestions, not to mention that the final responsibility of the decision should be on the care provider. In a previous study, diagnosis of middle ear disease by nonspecialists was reportedly only 30% in a study with primary care trainees [36] and 50% in a study with pediatricians just after finishing a continuing medical examination course [37]. Even for otolaryngologists, the accuracy of diagnosing otitis media using a pneumatic otoscope was 73% [37], which implies that accurate diagnosis using otoscopy is challenging [13,14,17]. Computer-aided diagnosis may be beneficial for both experts and nonotolaryngologists, for example, with our proposed ML model.

It is worth mentioning that our ML model acted as an otolaryngologist since the interrater variability (kappa) score between the ML model and otolaryngologist was similar to the kappa score between otolaryngologists (Figure 3B, ENT and ENT+ML'). Therefore, having our ML models interpret otoendoscopic images may be similar to having an on-demand otolaryngology consultant. Considering the shortage of specialists, nonotolaryngologists may combine our image interpretation results and clinical manifestations, which our ML does not consider, to deliver an accurate diagnosis and care for their patients.

Limitations

We point out the limitations and future directions of our study. Due to privacy issues, we could not perform our model outside the institution, and external validation could not be performed.

However, our otoendoscopic images were acquired from a diverse set of types of imaging equipment, which may mimic external validation. Also, as pointed out in our Methods section, one image may have multiple pathologies but was labeled according to the labeling priority. Multilabel classification should be conducted in the future, along with multimodal models that consider a patient's clinical information. Last but not least, although our ML models showed good accuracy in analyzing images, the current model does not consider additional clinical information, which most clinicians consider when making a diagnosis. Therefore, our ML model's higher accuracy in image translation may not necessarily correlate to better diagnostic expertise to physicians in the real world.

Comparison With Prior Work

In our previous study [9], we also classified ear disease into 6 entities but tested our model in a 5-fold cross-validation manner. Therefore, overall accuracy was less affected by classes of lower prevalence, showing inferior performance when applying the model in real-world clinics. A more recent study by Byun et al [38] assessed the effects of diagnostic assistant systems when used by otolaryngology residents. However, the diversity of disease was limited (only 4 diseases) and did not cover all ear diseases, especially external ear diseases and tumors. Also, the test set's size was small and did not test under various circumstances, that is balanced and imbalanced test sets. Our work addressed these effects and tests in both settings with a larger test set and more importantly, nonotolaryngologists, who may benefit most from using diagnostic assistance. We also measured the interrater reliability using kappa statistics, proving our proposed ML model similar to an otolaryngologist rather than a general practitioner.

Conclusions

Among the many potential differences, we focused on the data-driven classification bias of AI due to class imbalances of data in real-world clinics. ML is trained to find statistically optimal features from a large amount of training data in a way that improves the overall classification accuracy. Different numbers of samples in different classes due to imbalanced incidences inherent in the human population induce difficulty in building a reliable ML model. Based on the results of class imbalance, sample size, and accuracy (Figure 2), we still prefer a large but imbalanced dataset to a small but balanced dataset for a robust ML model. Therefore, our future system should analyze the strengths and weaknesses of the human experts and weigh the ML results to make suggestions depending on the situation: It provides strong suggestions when ML is superior and weak suggestions when ML is vulnerable. Along with suggestions, the system may display relative confidence in its diagnostic ability. Especially in atypical and rare diseases, this approach may provide more robust diagnoses, making the prediction system similar to consulting a fellow expert trained in a different institution for a second opinion.

Considering the practical situation in the clinical field that is short of otolaryngology specialists, clinicians may utilize our diagnostic assistance systems to deliver reliable patient care, while keeping in mind that the ML model does not consider additional clinical information and could be biased toward prevalent diseases.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NFR) grant funded by the Korea government (MSIP; 2020R1A2C3005787). We thank Dr. Young Min Moon, Department of Otorhinolaryngology, Yonsei University College of Medicine; Dr. Mi Jang, Department of Internal Medicine, Ilsan Paik Hospital, Inje University College of Medicine; Dr. Sunhee Kim, Department of Family Medicine; Soon Chun Hyang University College of Medicine; Dr. Joo Hyung Lee; Dr. Dong-Uk Lee; Dr. Mid-Eum Moon; and Dr. Jae-Min Choi (all private clinics) for participating in the evaluation of otoendoscopic images. In addition, this research was supported by the Brain Research Program through the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2017M3C7A1030750).

Conflicts of Interest

None declared.

Multimedia Appendix 1
Mixup strategy for oversampling and focal loss for loss function.
[DOCX File , 14 KB - medinform_v9i12e33049_app1.docx]

Multimedia Appendix 2
Effects of augmentation techniques applied to classification models.
[DOCX File , 40 KB - medinform_v9i12e33049_app2.docx]

Multimedia Appendix 3
Confusion matrix in imbalanced and balanced test set.
[DOCX File , 198 KB - medinform_v9i12e33049_app3.docx]

Multimedia Appendix 4
Supplementary tables.
[DOCX File , 67 KB - medinform_v9i12e33049_app4.docx]

References

1. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human-computer collaboration for skin cancer recognition. Nat Med 2020 Aug 22;26(8):1229-1234. [doi: 10.1038/s41591-020-0942-0] [Medline: 32572267]
2. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. Lancet Oncol 2019 Jul;20(7):938-947 [FREE Full text] [doi: 10.1016/S1470-2045(19)30333-X] [Medline: 31201137]
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 Jan 7;25(1):44-56. [doi: 10.1038/s41591-018-0300-7] [Medline: 30617339]
4. Mazo C, Bernal J, Trujillo M, Alegre E. Transfer learning for classification of cardiovascular tissues in histological images. Comput Methods Programs Biomed 2018 Oct;165:69-76. [doi: 10.1016/j.cmpb.2018.08.006] [Medline: 30337082]
5. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016 Dec 13;316(22):2402-2410. [doi: 10.1001/jama.2016.17216] [Medline: 27898976]
6. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Med 2018 Nov 27;15(11):e1002699 [FREE Full text] [doi: 10.1371/journal.pmed.1002699] [Medline: 30481176]
7. Slotnick HB. How doctors learn: physicians' self-directed learning episodes. Acad Med 1999 Oct;74(10):1106-1117. [doi: 10.1097/00001888-199910000-00014] [Medline: 10536633]
8. Stern E. Individual differences in the learning potential of human beings. NPJ Sci Learn 2017 Jan 12;2(1):2 [FREE Full text] [doi: 10.1038/s41539-016-0003-0] [Medline: 30631449]
9. Cha D, Pae C, Seong S, Choi J, Park HJ. Automated diagnosis of ear disease using ensemble deep learning with a big otoendoscopy image database. EBioMedicine 2019 Jul;45:606-614 [FREE Full text] [doi: 10.1016/j.ebiom.2019.06.050] [Medline: 31272902]
10. Chronic suppurative otitis media: burden of illness and management options. World Health Organization. 2004. URL: https://apps.who.int/iris/handle/10665/42941 [accessed 2021-11-09]
11. Multi-country assessment of national capacity to provide hearing care. World Health Organization. 2013. URL: https://www.who.int/publications/i/item/9789241506571 [accessed 2021-11-09]
12. Fagan J, Jacobs M. Survey of ENT services in Africa: need for a comprehensive intervention. Glob Health Action 2009 Mar 19;2(1):1932 [FREE Full text] [doi: 10.3402/gha.v2i0.1932] [Medline: 20027268]
13. Myburgh HC, van Zijl WH, Swanepoel D, Hellström S, Laurent C. Otitis media diagnosis for developing countries using tympanic membrane image-analysis. EBioMedicine 2016 Mar;5:156-160 [FREE Full text] [doi: 10.1016/j.ebiom.2016.02.017] [Medline: 27077122]
14. Moberly AC, Zhang M, Yu L, Gurcan M, Senaras C, Teknos TN, et al. Digital otoscopy versus microscopy: How correct and confident are ear experts in their diagnoses? J Telemed Telecare 2017 May 08;24(7):453-459. [doi: 10.1177/1357633x17708531]
15. Niermeyer WL, Philips RHW, Essig GF, Moberly AC. Diagnostic accuracy and confidence for otoscopy: Are medical students receiving sufficient training? Laryngoscope 2019 Aug;129(8):1891-1897. [doi: 10.1002/lary.27550] [Medline: 30329157]
16. Sanna M, Russo A, Caruso A, Taibah A, Piras G. Color Atlas of Endo-Otoscopy: Examination-Diagnosis-Treatment. New York, NY: Thieme; 2017.
17. Pichichero M, Poole M. Comparison of performance by otolaryngologists, pediatricians, and general practioners on an otoendoscopic diagnostic video examination. Int J Pediatr Otorhinolaryngol 2005 Mar;69(3):361-366 [FREE Full text] [doi: 10.1016/j.ijporl.2004.10.013] [Medline: 15733595]
18. ImageNet. URL: http://www.image-net.org [accessed 2021-11-09]
19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/CVPR.2016.90]
20. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.308]
21. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-ResNet and the impact of residual connections on learning. 2017 Presented at: Thirty-First AAAI Conference on Artificial Intelligence; February 4-9, 2017; San Francisco, CA.
22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Cornell University. 2015. URL: https://arxiv.org/abs/1409.1556 [accessed 2021-11-09]

23. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. 2018 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT. [doi: 10.1109/cvpr.2018.00745]
24. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); July 21-26, 2017; Honolulu, HI. [doi: 10.1109/cvpr.2017.243]
25. Liu C, Zoph B, Neumann M, Shlens J, Hua W, Li LJ, et al. Progressive Neural Architecture Search. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11205. Cham, Switzerland: Springer; 2018:19-35.
26. Zoph B, Vasudevan V, Shlens J, Le QV. Learning transferable architectures for scalable image recognition. 2018 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; June 18-23, 2018; Salt Lake City, UT. [doi: 10.1109/cvpr.2018.00907]
27. Kingma D, Ba J. Adam: A method for stochastic optimization. Cornell University. 2014. URL: https://arxiv.org/abs/1412.6980 [accessed 2021-11-09]
28. Masters D, Luschi C. Revisiting small batch training for deep neural networks. Cornell University. 2018. URL: https://arxiv.org/abs/1804.07612 [accessed 2021-11-09]
29. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G. Pytorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019. URL: https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html [accessed 2021-11-09]
30. Howard J, Gugger S. Fastai: a layered API for deep learning. Information 2020 Feb 16;11(2):108. [doi: 10.3390/info11020108]
31. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. mixup: Beyond empirical risk minimization. Cornell University. 2018. URL: https://arxiv.org/abs/1710.09412 [accessed 2021-11-09]
32. Lin TY, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell 2020 Feb;42(2):318-327. [doi: 10.1109/TPAMI.2018.2858826] [Medline: 30040631]
33. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 2016 Jul 02;33(3):613-619. [doi: 10.1177/001316447303300309]
34. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
35. Li D, Liu C, Hu S. A learning method for the class imbalance problem with medical data sets. Comput Biol Med 2010 May;40(5):509-518 [FREE Full text] [doi: 10.1016/j.compbiomed.2010.03.005] [Medline: 20347072]
36. Oyewumi M, Brandt MG, Carrillo B, Atkinson A, Iglar K, Forte V, et al. Objective evaluation of otoscopy skills among family and community medicine, pediatric, and otolaryngology residents. J Surg Educ 2016;73(1):129-135. [doi: 10.1016/j.jsurg.2015.07.011] [Medline: 26364889]
37. Pichichero ME, Poole MD. Assessing diagnostic accuracy and tympanocentesis skills in the management of otitis media. Arch Pediatr Adolesc Med 2001 Oct;155(10):1137-1142. [doi: 10.1001/archpedi.155.10.1137] [Medline: 11576009]
38. Byun H, Yu S, Oh J, Bae J, Yoon MS, Lee SH, et al. An assistive role of a machine learning network in diagnosis of middle ear diseases. J Clin Med 2021 Jul 21;10(15):3198 [FREE Full text] [doi: 10.3390/jcm10153198] [Medline: 34361982]

Abbreviations

ADAM: Adaptive Moment Estimation
AI: artificial intelligence
CNN: convolutional neural network
DL: deep learning
EAC: external auditory canal
HI: human intelligence
ML: machine learning

Edited by G Eysenbach; submitted 22.08.21; peer-reviewed by J Bernal; comments to author 24.09.21; revised version received 29.09.21; accepted 12.10.21; published 08.12.21.

Please cite as:
Cha D, Pae C, Lee SA, Na G, Hur YK, Lee HY, Cho AR, Cho YJ, Han SG, Kim SH, Choi JY, Park HJ
Differential Biases and Variabilities of Deep Learning–Based Artificial Intelligence and Human Experts in Clinical Diagnosis: Retrospective Cohort and Survey Study
JMIR Med Inform 2021;9(12):e33049
URL: https://medinform.jmir.org/2021/12/e33049
doi:10.2196/33049
PMID:34889764

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org