JMIR Medical Informatics

Clinical informatics, decision support for health professionals, electronic health records, and eHealth infrastructures
Volume 9 (2021), Issue 9   ISSN: 2291-9694

Contents

Viewpoints

Original Papers

Reviews

Corrigenda and Addenda

Viewpoint

Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice

Viraj Kulkarni[1*], MSc; Manish Gawali[1*], BE; Amit Kharat[1,2*], DPhil

[1]DeepTek Inc, Pune, India

[2]D Y Patil University, Pune, India

[*]all authors contributed equally

Corresponding Author:
Manish Gawali, BE
DeepTek Inc
2rd Floor, Alacrity Innovation Centre
3, Baner Rd, Pallod Farms, Baner
Pune, 411045
India
Phone: 91 72760 60080
Email: manish.gawali@deeptek.ai

Abstract

The use of machine learning to develop intelligent software tools for the interpretation of radiology images has gained widespread attention in recent years. The development, deployment, and eventual adoption of these models in clinical practice, however, remains fraught with challenges. In this paper, we propose a list of key considerations that machine learning researchers must recognize and address to make their models accurate, robust, and usable in practice. We discuss insufficient training data, decentralized data sets, high cost of annotations, ambiguous ground truth, imbalance in class representation, asymmetric misclassification costs, relevant performance metrics, generalization of models to unseen data sets, model decay, adversarial attacks, explainability, fairness and bias, and clinical validation. We describe each consideration and identify the techniques used to address it. Although these techniques have been discussed in prior research, by freshly examining them in the context of medical imaging and compiling them in the form of a laundry list, we hope to make them more accessible to researchers, software developers, radiologists, and other stakeholders.

(JMIR Med Inform 2021;9(9):e28776) doi:10.2196/28776

KEYWORDS

artificial intelligence; AI; machine learning; deep learning; radiology; privacy; neural networks; deployment

Introduction

Although radiology imaging has emerged as an indispensable tool in diagnostic medicine, there is a worldwide shortage of qualified radiologists to read, interpret, and report these images [1,2]. The volume of images is growing faster than the number of radiologists. The high workload that this causes leads to errors in diagnosis because of human fatigue, unacceptable delays in reporting, and stress and burnout in radiologists. On the other hand, artificial intelligence (AI) and machine learning models have shown remarkable performance in the automated evaluation of medical images [3-5]. In this situation, hospitals are increasingly drawn toward adopting computer-aided detection technologies for processing scans. These technologies show considerable promise in improving diagnostic accuracy, reducing reporting time, and boosting radiologist productivity.

Supervised machine learning, the most common form of machine learning, works in two phases. In the first phase, the algorithm implemented as a software reads a training data set consisting of images along with their corresponding labels. It processes these data, extracts patterns from it, and learns a function that maps an input image to its corresponding label. The learned mapping function along with the extracted patterns are mathematically represented in the form of the trained model. This is called the *training* phase. In the second phase, called the *inference* phase, the trained model is used to read input images and make predictions. Artificial neural networks are a class of machine learning algorithms; artificial neural networks with many layers are called deep neural networks. In the literature, the terms deep learning, AI, and artificial neural networks tend to be used interchangeably. In this paper, we use *machine learning* to broadly refer to all the terms mentioned earlier in addition to conventional machine learning algorithms, such as linear regression, support vector machines, decision trees, and random forests.

The development of machine learning models for radiology involves many challenges. High-quality training data are vital for good model performance [6] but are difficult to obtain. Available data may lack volume or diversity. It may be scattered across multiple hospitals. Even if the image data are available, they may not be labeled. Radiology scans suffer from a high degree of interreader variability, where 2 or more radiologists label the data inconsistently [7,8]; this may lead to noise or uncertainty in the ground truth labels. The distribution of target classes may be heavily skewed, especially for rare pathologies. This imbalance in class representation is often accompanied by unequal misclassification costs across classes. Care must be taken when dealing with imbalanced data sets, and this sometimes requires using special performance measures [9]. A model that works well on data from one hospital may perform poorly on data from a different hospital [10]. Similarly, a model deployed in practice at a hospital may experience a gradual decay in performance at the same hospital [11]. Machine learning models have been shown to be vulnerable to malicious exploits and attacks [12-14]. To support adoption by radiologists, the deployed models should be able to explain their decisions [15], and they should not discriminate patients on the basis of gender, ethnicity, age, income, among others [16].

This study has a simple structure. In the Key Considerations section, we enumerate the key considerations that machine learning researchers should acknowledge and address. For each consideration, we describe the common challenges and their significance before suggesting solutions to overcome them. In the Conclusions section, we discuss other overarching limitations that hinder the adoption of machine learning in clinical radiology practice.

Key Considerations

Insufficient Training Data

Machine learning models are data hungry, and their performance depends heavily on the characteristics of the data used to train them [6]. The training set size has a direct and significant effect on the performance of the models. However, the heterogeneity and diversity of the training data influence the ability of the models to generalize to unseen data sources [17]. To develop robust machine learning models, researchers need access to large medical data sets that adequately represent data diversity in terms of population features such as age, gender, ethnicity, and medical conditions and imaging features such as equipment manufacturers, image capture settings, and patient posture. Most available data sets in medical imaging do not meet these requirements [18-20]. As many critical conditions have low rates of occurrence, very little data are available for them. Machine learning models trained using these scanty data to diagnose rare conditions fail to perform well in practice even if they demonstrate good performance in retrospective evaluations.

Several methods have been proposed for dealing with insufficient data for training models. Data augmentation techniques including geometric transformations and color-space transformations can enhance the quantity and variety of training data [21]. Generative adversarial networks have shown success in generating synthetic images for rare pathologies, which can be further used for model training [22]. Although these techniques allow models to be trained on scarce data by artificially increasing the variation in the data set, they cannot serve as a substitute for high-quality data.

Decentralized Data Sets

Many medical data sets are naturally distributed across multiple storage devices connected to networks owned by different institutions. In traditional machine learning settings, these data sets need to be consolidated into a single repository before training the models. Moving large volumes of data across networks poses several logistical and legal challenges [23]. Government policies such as the General Data Protection Regulation [24], the Health Insurance Portability and Accountability Act [25], and the Singapore Personal Data Protection Act [26] also stipulate restrictions on sharing and movement of data across national borders.

Privacy-preserving distributed learning techniques such as federated learning [27] and split learning [28] enable machine learning models to train on decentralized data sets at multiple client sites without moving the data and compromising privacy. Implementing these techniques, however, entails additional overheads, which may render the exercise unfeasible. These overheads include the high cost of developing software that supports these technologies, the high network communication bandwidth required, the orchestration effort in deploying it at multiple sites, and the possibly reduced performance of the predictive models [29]. Federated learning generates a global shared model for all clients, leading to situations where, for some clients, the local models trained on their private data perform better than the global shared model. In such situations, additional personalization techniques may be required to fine-tune the global model individually for each client [30].

High Cost of Annotations

Supervised machine learning requires the annotation of radiology images before they can be used to train the model. Image-level annotations classify each image into one or more classes, whereas region-level annotations highlight regions within an image and classify each region into one or more classes. As the predictive performance of the model is directly influenced by the quality of the annotations, it is imperative that the data are annotated by qualified radiologists or medical practitioners [31]. This makes the process of annotation exorbitantly expensive in many cases.

Several efforts have been made to use natural language processing (NLP) techniques to automatically annotate images by extracting labels from radiology text reports [32-34]. Semisupervised approaches can be used when a small amount of labeled data are available along with a larger amount of unlabeled data [35,36]. As manual annotations are expensive, AI-based automated image annotation techniques can be considered [37].

Ambiguous Ground Truth

As hospital data sets usually contain images accompanied by their text reports, many projects are kickstarted by using NLP techniques to automatically annotate the images using the reports. Radiology reports, however, vary widely in their comprehensiveness, style, language, and format [38]. Even if state-of-the-art NLP manages to accurately extract all the findings from the text report, the report itself may not mention all the findings. Olatunji et al [39] showed that there is a large discrepancy between what radiologists see in an image and what they mention in the report; reporting radiologists usually document only those findings that are relevant to the immediate clinical context and are likely to miss reporting nonactionable or borderline findings.

Radiology images suffer from significant interreader variability, where 2 or more experts may disagree on the findings from a scan [7,8,40-43]. Sakurada et al [44], for instance, report low interreader κ values ranging from 0.24 to 0.63 for assessment of different pathologies from chest radiographs. In practice, annotation workflows generally engage a single reader to assign ground truth labels to images. An improvement over this involves engaging multiple independent readers and considering their majority vote as the ground truth label. However, single reader or majority vote approaches may miss labeling challenging but critical findings.

This risk can be mitigated by using multiphase reviews [45] or expert adjudication [46] to create high-quality labels. Majkowska et al [46] showed that adjudication improved the consensus among radiologists to 96.8% compared with 41.8% after the first independent readings when assessing chest radiographs. Raykar et al [47] proposed a probabilistic approach to determine the *hidden* ground truth from labels assigned by multiple radiologists and demonstrated that this method is superior to majority voting. In some clinical settings, radiology imaging is used for initial screening before conducting subsequent confirmatory tests. For example, chest x-ray scans may be used as a first-line test before subsequently conducting a computed tomography scan, laboratory tests, or biopsy. Data from these subsequent tests, if available, should be used to validate and correct the labels assigned to the images from the screening test. In situations where human-labeled ground truth is noisy or ambiguous, developing a process to reduce variability and improve label quality may yield better models than attempts to improve model performance on the original labels by other means.

Imbalance in Class Representation

Class imbalance occurs when all label classes are not equally represented in the training data set [48]. This is a common situation when building binary classifiers for medical data sets where the number of *normal* examples in which the target abnormality is absent is many times larger than the number of *abnormal* examples in which it is present. As machine learning models are usually trained by optimizing a loss function across all training examples, the trained models tend to favor the majority class over the minority class. Researchers have empirically evaluated the adverse effect of class imbalance on classification performance in several studies [9,49-53].

Class imbalance can be handled at the data level or the algorithmic level. Resampling strategies can be used to address imbalances in the training data by either undersampling the majority classes or oversampling the minority classes. Many comparative evaluations of these approaches exist, sometimes with contradictory conclusions. Drummond et al [54], for instance, argued that undersampling works better than oversampling, whereas Batista et al [55] reported superior performance using oversampling. However, we caution the reader against hasty generalizations and note that these comparisons are highly dependent on the data set, the machine learning algorithm, the sampling technique used, and the parameters of the experiments. Chawla et al [56] proposed the synthetic minority oversampling technique, a technique to generate synthetic examples to balance the data set, and showed that the combination of synthetic minority oversampling technique and undersampling performs better than plain undersampling or oversampling. Similarly, oversampling can also be performed using geometric augmentations, color-space augmentations, or generative models to produce synthetic images. An imbalance in the number of examples can also be addressed at the algorithmic level using methods such as one-class classification, outlier or anomaly detection, regularized ensembles, and custom loss functions [9,57-60].

Asymmetric Misclassification Costs

Standard machine learning settings assume that all misclassifications between classes are equal and incur the same penalty. This assumption is not true for many medical imaging problems. For example, the cost of classifying a *normal* scan as *abnormal* may be very different from the cost of classifying an *abnormal* scan as *normal*.

This asymmetrical nature of the classification problem can be handled either at the time of deployment or during development. The trained model can be tuned to achieve higher sensitivity or specificity according to the requirements at deployment time. Alternatively, the variation in misclassification penalties can be represented as a cost matrix, where each element C(i,j) represents the penalty of misclassifying an example of class i as class j. The model can then be trained by minimizing the overall cost as defined by the asymmetrical loss function. For more details, we refer the reader to the literature on cost-sensitive learning [9,61,62].

Relevant Performance Measures

Machine learning researchers and practitioners tend to ignore the question of how model performance should be evaluated in cases of imbalanced data sets and asymmetric misclassification costs. Most binary classification models produce a continuous-valued output score. This score is converted into discrete binary labels using a cutoff threshold. Owing to its simplicity, it is tempting to use accuracy, defined as the percentage of predictions that are correct, as a measure of performance. However, in the case of imbalanced data sets, accuracy is ineffective and provides an incomplete and often misleading picture of the ability of the classifier to discriminate between the two classes [63,64].

Using two or more measures such as sensitivity, specificity, and precision, provides a better picture of the discriminative performance of a classifier [65]. However, these measures depend on the cutoff threshold mentioned earlier. Furthermore, the decision to set the threshold is often guided not by technology but by business or domain concerns. Comparing the two models by considering multiple performance measures across different operating thresholds is challenging. The receiver operating characteristic curve, on the other hand, captures the model performance at all threshold operating points. The area under the receiver operating characteristic curve (AUROC) thus serves as a single numerical score that represents the performance of the model across all operating threshold points. This has made AUROC a metric of choice for reporting the classification performance of machine learning models. Unfortunately, AUROC too can be deceptive when dealing with imbalanced data sets and may provide an overly optimistic view of performance [9]. The precision-recall curve and the area under it are more suitable for describing classification performance when data sets are imbalanced [66,67]. Drummond and Holte proposed cost curves that describe the classification performance over asymmetric misclassification costs and class distributions [68,69]. Table 1 shows how accuracy can be misleading because of imbalanced data sets.

Table 1. Example illustrating how accuracy can be misleading in case of imbalanced data sets.

	Predicted as negative	Predicted as positive	Total
Actual negative	80	10	90
Actual positive	5	5	10
Total	85	15	100

In the confusion matrix mentioned earlier, of 100 test examples, 90 are negative and 10 are positive. The classifier predicts 85 of them as negative and 15 as positive. This gives a high accuracy of 0.85 and a high specificity of 0.89. However, the complete picture is seen when we consider the low sensitivity of 0.50 and precision of 0.33.

Generalization of Models to Unseen Data Sets

Machine learning models are routinely evaluated on a hold-out set taken from the same source as the training set [70]. The available data are divided into two parts. One part is used to train and validate the models. The second part, called the test set or hold-out set, is used to estimate the final performance of the trained model when deployed. The underlying premise is that the data used to train the model are representative of the data that the model will encounter during clinical use. This assumption is often violated in practice, and this makes the performance on the hold-out set an unreliable indicator of future performance in clinical deployment.

Poor generalization of models to diverse patient groups is one of the biggest hurdles for the adoption of AI and machine learning in health care. One reason for the poor generalization is the difference in the image characteristics between images from the training sites and those from the deployment site. This variation, also known as data set shift, can occur because of differences in hospital procedures, equipment manufacturers, image acquisition parameters, disease manifestations, patient populations, among others. Owing to the data set shift, models trained using data from one hospital may perform poorly on data from another hospital [71]. We note here that this inability to generalize to data sets from an unseen origin is different from the problem of overfitting, where the model shows poor performance even on test sets from the same origin. Learning irrelevant confounders instead of relevant features is another reason why models fail to generalize to data from unseen origins. Machine learning models are notorious for exploiting confounders in the training data. For example, Zech et al [72] showed that a pneumonia classification model trained on data from 2 hospitals learned to leverage the difference between prevalence rates at the 2 hospitals instead of the relevant visual features.

Data augmentation can improve model generalization by increasing the variations in the training set [73]. Image processing techniques, including standardization, normalization, reorientation, registration, and histogram matching, can be used to harmonize images sourced from different origins and remove domain bias. However, Glocker et al [74] showed that even with a state-of-the-art image preprocessing pipeline, these techniques for harmonization were unable to remove scanner-specific bias, and machine learning models were easily able to discriminate between the different origins of the data.

Domain adaptation techniques can be used to fine-tune models to a new target domain by narrowing the gap between the source and target domains in a domain-invariant feature space [75-79]. On the other hand, domain generalization techniques attempt to train models that are sensitive only to features relevant to the classification task but insensitive to confounding features that differentiate between the domains [80-85].

Model Decay

Model decay refers to the phenomenon in which the performance of a deployed machine learning model deteriorates over time [11]. Supervised machine learning algorithms extract patterns from the training data to learn a mapping between independent input variables and a dependent target variable. This process involves making an implicit assumption that the data encountered in deployment will be stationary and will not change over time; this assumption is often violated in practice because of the changes in hospital workflows, imaging equipment, patient groups, evolving adoption of AI solutions, among others.

Model decay occurs owing to changes in the underlying data. These changes can be broadly classified into three types: (1) *covariate shift* occurs when there are changes in the distribution of the independent input variables (eg, the average age of the population increases over time); (2) *prior probability shift* occurs when there are changes in the distribution of the dependent target variables (eg, the prevalence of a particular disease in the target population may change because of seasonality or an epidemic); and (3) *concept drift* occurs when there are changes in the relationship between the independent and dependent variables (eg, changes in a hospital's diagnostic protocols or a radiologist's interpretation regarding which visual manifestations should or should not be considered indicative of a pathology). These changes can be sudden, gradual, or cyclic.

Detecting model decay requires continuous monitoring of the deployment time performance against a human-labeled subsample of the data. If the performance drops below a predetermined threshold, an alarm is triggered, and the model is retrained or fine-tuned using the most recent data. This retraining can also be conducted periodically as a routine maintenance activity. For more details, including theoretical frameworks for understanding model decay or practical solutions, readers can refer to additional reviews [11,86-89].

Adversarial Attacks

An adversarial example is constructed by deliberately injecting perturbations in the original image to trick the model into misclassifying the label for that image [12]. Machine learning models are susceptible to manipulation using such adversarial examples [90,91]. Data poisoning attacks [13] introduced adversarial examples in the training data to manipulate the diagnosis of the model being developed. On the other hand, evasion attacks [14] use adversarial examples to influence predictions during deployment. Health care is a huge economy, and many decisions regarding diagnosis, reimbursements, and insurance may be governed or assisted by algorithms in the near future. Hence, the discovery of these vulnerabilities has raised pressing concerns regarding the safety and usability of machine learning models in clinical practice.

Qayyum et al [92] provided a detailed taxonomy of defensive techniques against adversarial attacks by grouping them into three broad categories: (1) reconstructing the training or testing data to make it more difficult to manipulate [90,93-96], (2) modifying the model to make it more resilient to adversarial examples [97-101], and (3) using auxiliary models or ensembles to detect and neutralize adversarial examples [102-106]. Adversarial attacks and their countermeasures are an evolving research area, and there are excellent reviews for the same [107-109].

Explainability

The power of neural networks to uncover hidden relationships between variables and use them to make predictions is tempered by one disadvantage: the exact process the neural network uses to arrive at a decision is unclear to humans. This is why neural networks are sometimes called black boxes whose inner workings cannot be observed. To what extent can we delegate decision-making to machines while we remain unaware of how the machine arrives at a decision is a key question that stands in the way of adopting algorithms in many industries, including autonomous vehicles, law, finance, among others.

Algorithmic explainability is especially important in medicine, where stakes are high, and the field is prone to litigation. In the context of radiology, explainability can be improved by using localization models that highlight the region of interest within the scan that is suspected to contain the abnormality instead of classification models that only indicate the presence or absence of an abnormality. However, the development of localization models also requires training data to have region-based annotations in the form of bounding boxes or free-form masks. Where region-based annotations are not available, saliency maps [110] and explainability frameworks [111] can be used to identify a region within the image that most contributes to a particular decision. Another way to improve the user's trust in the models is to predict a confidence score in addition to the prediction. For example, instead of merely stating the prediction "Probability of Tuberculosis: 75%," the system should also state the model's confidence "Probability of Tuberculosis: 75%, Confidence in this prediction: Low." Deployment settings where predictive models are used to autonomously make decisions demand more stringent conditions of explainability than settings where the models are used to guide humans who make the final decisions. A comprehensive analysis of explainers in the domain of computer vision was performed by Buhrmester et al [112].

There have been calls to limit the use of AI and machine learning only to rule-based systems in fields where algorithmic decisions affect human lives [113]. These systems are transparent and can trace the relationship between the input and the output as a sequence of rules that humans can understand. We find two problems with this approach. First, one of the chief advantages of using neural networks is that they can model complex relationships that *humans cannot understand*, and this is precisely what makes them so effective. Second, making decision systems transparent and explainable also makes them vulnerable to malicious attacks. A transparent rule-based method to make decisions can be *hacked*, *gamed*, or exploited more easily than a black box system [114,115].

Fairness and Bias

Algorithmic systems play a key role in guiding decisions that impact the delivery of health care to patients. Therefore, it is desirable that these systems are free of societal biases and their decisions are fair and equitable. Unfortunately, many existing data sets [18,43] reflect the biases of the societies that they represent [116], and it is difficult to detect and remove bias inherent in the training data. Obermeyer et al [16] showed, for example, that a widely used algorithmic system exhibited racial bias against Black patients, which reduced the number of Black patients eligible for extra care by more than half.

In principle, a predictive model is considered fair if it does not discriminate patients on the basis of sensitive variables such as gender, ethnicity, disability, and income. However, translating this seemingly simple principle into practice is a challenging issue. Researchers have developed numerous mathematical definitions of fairness and techniques to implement them [117]. One technique, for example, excludes sensitive variables from the input when training the model. Another technique is to tune the model so that it demonstrates the same level of performance as measured by sensitivity, specificity, among others, across all groups defined by the sensitive variables. Corbett-Devies and Goel [118] show that although appealing, these techniques suffer from significant statistical limitations and may adversely affect the same groups they were designed to protect. Pleiss et al [119] show how different definitions of fairness can be mutually incompatible, and a model designed to comply with one definition may violate another equally valid definition.

Algorithmic bias and fairness are evolving fields of research that lie at the intersection of machine learning, public policy, law, and ethics. We believe that fairness is not inherently a technological problem but a societal one. Coercing technology to solve it can lead to automated systems that tick the right boxes for some arbitrary definition of fairness but eventually end up worsening social inequality and discrimination behind a veneer of technical neutrality [120].

Clinical Validation

A comprehensive evaluation to assess the predictive performance and clinical utility of a model must be conducted before it can be deployed in clinical practice. When a model is evaluated on a hold-out set collected from the same sources from which the training data are collected, the evaluation is called an *internal* validation. When a data set from an *unseen* source is used to evaluate the model, the evaluation is called *external* validation. As described earlier in the section *Generalization of Models to Unseen Data Sets*, the lack of generalization to unseen data sources is one of the biggest challenges in the adoption of machine learning in practice. Despite this, only a fraction of the published studies report the results of an external validation [121]. Mahajan et al [122] presented examples to advocate the case for independent external validation of models before deployment and described a framework for the same. Park et al [31] proposed a methodology with a checklist for evaluating the clinical performance of the models. The TRIPOD statement [123] provides guidelines for transparent reporting of the development and validation of prediction models for prognosis and diagnosis models. Although retrospective evaluations allow machine learning developers to test their models on large and diverse data sets, prospective evaluations allow testing in real-world environments; both types of evaluations are equally important and should be meticulously carried out before full-scale adoption.

Conclusions

We identify the key challenges that researchers face in developing accurate, robust, and usable machine learning models that can create value in clinical radiology practice. These challenges and the techniques to overcome them have been discussed previously in a piecemeal manner in prior research literature. In this study, we re-examined them in the context of medical imaging. By compiling them in the form of a laundry list, we hope to make this research more readily accessible.

Hospital workflows and practices vary widely from one hospital to another, even within the same geography. This increases the difficulty of seamlessly integrating predictive models into hospital workflows. The nonuniformity in workflows also raises the question of whether the reported performance of a model is reproducible in a different clinical context. This is an ongoing research, and satisfactory solutions are yet to be found.

The ultimate objective of diagnostic machine learning models is to improve patient outcomes. However, improvement in diagnostic performance does not, by itself, cause an improvement in patient outcomes [31]. Radiological diagnosis is only one of the many steps that eventually leads to treatment. Therefore, a computerized diagnostic system must be placed appropriately in the workflow. How the system presents the results to the reporting radiologist and what action the radiologist takes on receiving them are important factors that influence the usefulness of the system in practice.

On the one hand, medical imaging is a broad and complex field that encompasses numerous imaging modalities, pathological conditions, and diagnostic protocols. On the other hand, machine learning is an active area of research with thousands of new techniques published every year. The combined diversity of both fields along with nonuniform hospital practices, regulatory restrictions on data sharing, and lack of standardized reporting of results make it difficult to clearly assess the role and potential of machine learning applications in medical imaging. We believe that machine learning has great potential in improving diagnostic accuracy, lowering reporting times, reducing radiologist workloads, and ultimately improving the delivery of health care. To realize this potential, however, a concerted across-the-board effort will be required from physicians, radiologists, patients, hospital administrators, data scientists, software developers, and other stakeholders.

Authors' Contributions

VK is the primary author of the paper and was responsible for conceiving the topic, reviewing the research literature, and writing the manuscript. MG assisted in surveying the research literature and writing some sections of the manuscript. AK is a senior radiologist who validated the manuscript from a clinical radiology perspective and assisted in editing the manuscript.

Conflicts of Interest

None declared.

References

1. Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. Br Med J 2017 Oct 11;359:j4683. [doi: 10.1136/bmj.j4683] [Medline: 29021184]

2. Nakajima Y, Yamada K, Imamura K, Kobayashi K. Radiologist supply and workload: international comparison--Working Group of Japanese College of Radiology. Radiat Med 2008 Oct;26(8):455-465. [doi: 10.1007/s11604-008-0259-2] [Medline: 18975046]
3. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJ. Artificial intelligence in radiology. Nat Rev Cancer 2018 Aug;18(8):500-510 [FREE Full text] [doi: 10.1038/s41568-018-0016-5] [Medline: 29777175]
4. Maretíc Z. Cnidarismus nudorum: A new epidemiological and clinical entity. Dermatologica 1986;172(2):123-125. [Medline: 2868931]
5. Shen D, Wu G, Suk H. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19:221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
6. Foody G, McCulloch MB, Yates WB. The effect of training set size and composition on artificial neural network classification. Int J Remote Sens 2007 May 03;16(9):1707-1723. [doi: 10.1080/01431169508954507]
7. Kerlikowske K, Grady D, Barclay J, Frankel SD, Ominsky SH, Sickles EA, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998 Dec 02;90(23):1801-1809. [doi: 10.1093/jnci/90.23.1801] [Medline: 9839520]
8. Moifo B, Pefura-Yone EW, Nguefack-Tsague G, Gharingam ML, Tapouh JR, Kengne A, et al. Inter-observer variability in the detection and interpretation of chest x-ray anomalies in adults in an endemic tuberculosis area. O J Med Imaging 2015 Sep;05(03):143-149. [doi: 10.4236/ojmi.2015.53018]
9. He H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng 2009 Sep;21(9):1263-1284. [doi: 10.1109/TKDE.2008.239]
10. Pooch E, Ballester P, Barros R. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv.org. 2020. URL: http://arxiv.org/abs/1909.01940 [accessed 2021-08-16]
11. Widmer G, Kubat M. Learning in the presence of concept drift and hidden contexts. Mach Learn 1996 Apr;23(1):69-101. [doi: 10.1007/bf00116900]
12. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. arXiv.org. 2014. URL: https://arxiv.org/abs/1312.6199 [accessed 2021-08-16]
13. Steinhardt J, Koh P, Liang P. Certified defenses for data poisoning attacks. arXiv.org. 2017. URL: https://arxiv.org/abs/1706.03691 [accessed 2021-08-16]
14. Biggio B, Corona I, Maiorca D. Evasion attacks against machine learning at test time. In: Machine Learning and Knowledge Discovery in Databases. Berlin, Heidelberg: Springer; 2013.
15. Brunese L, Mercaldo F, Reginelli A, Santone A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. Comput Methods Programs Biomed 2020 Nov;196:105608 [FREE Full text] [doi: 10.1016/j.cmpb.2020.105608] [Medline: 32599338]
16. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019 Oct 25;366(6464):447-453. [doi: 10.1126/science.aax2342] [Medline: 31649194]
17. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a radiologist's guide. Radiology 2019 Mar;290(3):590-606. [doi: 10.1148/radiol.2018180547] [Medline: 30694159]
18. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers R. ChestX-Ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI, USA. [doi: 10.1109/cvpr.2017.369]
19. Bustos A, Pertusa A, Salinas J, de la Iglesia-Vayá M. PadChest: a large chest x-ray image dataset with multi-label annotated reports. Med Image Anal 2020 Dec;66:101797. [doi: 10.1016/j.media.2020.101797] [Medline: 32877839]
20. Johnson A, Pollard T, Greenbaum N, Lungren M, Deng C, Peng Y, et al. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv.org. 2019. URL: http://arxiv.org/abs/1901.07042 [accessed 2021-08-16]
21. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data 2019 Jul 6;6(1):60. [doi: 10.1186/s40537-019-0197-0]
22. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: A review. Med Image Anal 2019 Dec;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
23. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health 2014 Nov 05;14:1144 [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]
24. Voigt P, von dem Bussche A. The EU General Data Protection Regulation (GDPR). Switzerland: Springer; 2017.
25. Annas GJ. HIPAA regulations - a new era of medical-record privacy? N Engl J Med 2003 Apr 10;348(15):1486-1490. [doi: 10.1056/NEJMlim035027] [Medline: 12686707]
26. Chik WB. The Singapore Personal Data Protection Act and an assessment of future trends in data privacy reform. Comput Law Secur Rev 2013 Oct;29(5):554-575. [doi: 10.1016/j.clsr.2013.07.010]
27. McMahan H, Moore E, Ramage D, Hampson S, Arcas B. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. 2017

Presented at: 20th International Conference on Artificial Intelligence and Statistics; Fort Lauderdale, Florida, USA; May 9-11, 2017 URL: http://proceedings.mlr.press/v54/mcmahan17a.html

28. Vepakomma P, Gupta O, Swedish T, Raskar R. Split learning for health: Distributed deep learning without sharing raw patient data. arXiv.org. 2018. URL: http://arxiv.org/abs/1812.00564 [accessed 2021-08-16]
29. Gawali M, Suryavanshi S, CS A, Madaan H, Gaikwad A, Bhanu Prakash KN, et al. Comparison of privacy-preserving distributed deep learning methods in healthcare. In: Proceedings of the Annual Conference on Medical Image Understanding and Analysis. 2021 Presented at: Annual Conference on Medical Image Understanding and Analysis; July 12-14, 2021; Oxford, United Kingdom. [doi: 10.1007/978-3-030-80432-9_34]
30. Kulkarni V, Kulkarni M, Pant A. Survey of personalization techniques for federated learning. In: Proceedings of the Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). 2020 Presented at: Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4); July 27-28, 2020; London, UK. [doi: 10.1109/worlds450073.2020.9210355]
31. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. Radiology 2018 Mar;286(3):800-809. [doi: 10.1148/radiol.2017171920] [Medline: 29309734]
32. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural language-based machine learning models for the annotation of clinical radiology reports. Radiology 2018 May;287(2):570-580. [doi: 10.1148/radiol.2018171093] [Medline: 29381109]
33. Smit A, Jain S, Rajpurkar P, Pareek A, Ng A, Lungren M. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020 Presented at: Conference on Empirical Methods in Natural Language Processing (EMNLP); Nov 16-20, 2020; Punta Cana URL: https://aclanthology.org/2020.emnlp-main.117.pdf [doi: 10.18653/v1/2020.emnlp-main.117]
34. Pons E, Braun LM, Hunink MG, Kors JA. Natural language processing in radiology: a systematic review. Radiology 2016 May;279(2):329-343. [doi: 10.1148/radiol.16142770] [Medline: 27089187]
35. Cheplygina V, de Bruijne M, Pluim JP. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med Image Anal 2019 May;54:280-296. [doi: 10.1016/j.media.2019.03.009] [Medline: 30959445]
36. Feyjie A, Azad R, Pedersoli M, Kauffman C, Ayed I, Dolz J. Semi-supervised few-shot learning for medical image segmentation. arXiv.org. 2020. URL: http://arxiv.org/abs/2003.08462 [accessed 2021-08-16]
37. Cheng Q, Zhang Q, Fu P, Tu C, Li S. A survey and analysis on automatic image annotation. Pattern Recognit 2018 Jul;79:242-259. [doi: 10.1016/j.patcog.2018.02.017]
38. Brady AP. Radiology reporting-from Hemingway to HAL? Insights Imaging 2018 Apr;9(2):237-246 [FREE Full text] [doi: 10.1007/s13244-018-0596-3] [Medline: 29541954]
39. Olatunji T, Yao L, Covington B, Rhodes A, Upton A. Caveats in generating medical imaging labels from radiology reports. arXiv.org. 2019. URL: http://arxiv.org/abs/1905.02283 [accessed 2021-08-16]
40. Rosenkrantz AB, Duszak R, Babb JS, Glover M, Kang SK. Discrepancy rates and clinical impact of imaging secondary interpretations: a systematic review and meta-analysis. J Am Coll Radiol 2018 Sep;15(9):1222-1231. [doi: 10.1016/j.jacr.2018.05.037] [Medline: 30031614]
41. Brouwer CL, Steenbakkers RJ, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D variation in delineation of head and neck organs at risk. Radiat Oncol 2012 Mar 13;7:32 [FREE Full text] [doi: 10.1186/1748-717X-7-32] [Medline: 22414264]
42. Njeh CF. Tumor delineation: the weakest link in the search for accuracy in radiotherapy. J Med Phys 2008 Oct;33(4):136-140 [FREE Full text] [doi: 10.4103/0971-6203.44472] [Medline: 19893706]
43. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. AAAI-19 2019 Jul 17;33(1):590-597. [doi: 10.1609/aaai.v33i01.3301590]
44. Sakurada S, Hang NT, Ishizuka N, Toyota E, Hung LD, Chuc PT, et al. Inter-rater agreement in the assessment of abnormal chest X-ray findings for tuberculosis between two Asian countries. BMC Infect Dis 2012 Feb 01;12:31 [FREE Full text] [doi: 10.1186/1471-2334-12-31] [Medline: 22296612]
45. Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. Med Phys 2011 Feb;38(2):915-931 [FREE Full text] [doi: 10.1118/1.3528204] [Medline: 21452728]
46. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology 2020 Feb;294(2):421-431. [doi: 10.1148/radiol.2019191293] [Medline: 31793848]
47. Raykar V, Yu S, Zhao L, Jerebko A, Florin C, Valadez G, et al. Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th Annual International Conference on Machine Learning. 2009 Presented at: ICML '09: 26th Annual International Conference on Machine Learning; Jun 14-18, 2009; Montreal Quebec Canada. [doi: 10.1145/1553374.1553488]

48. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019 Mar 19;6(1):27. [doi: 10.1186/s40537-019-0192-5]
49. Liu Y, Yu X, Huang JX, An A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Inf Process Manag 2011 Jul;47(4):617-631. [doi: 10.1016/j.ipm.2010.11.007]
50. Kim J, Kim J. The impact of imbalanced training data on machine learning for author name disambiguation. Scientometrics 2018 Jul;117(3-4):511-526. [doi: 10.1007/s11192-018-2865-9]
51. Chen H, Xiong F, Wu D, Zheng L, Peng A, Hong X, et al. Assessing impacts of data volume and data set balance in using deep learning approach to human activity recognition. In: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017 Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); Nov 13-16, 2017; Kansas City, MO, USA. [doi: 10.1109/bibm.2017.8217821]
52. Chawla N. Data mining for imbalanced datasets: an overview. In: Data Mining and Knowledge Discovery Handbook. Boston, MA: Springer; 2005:853-867.
53. Luque A, Carrasco A, Martín A, de las Heras A. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognit 2019 Jul;91:216-231. [doi: 10.1016/j.patcog.2019.02.023]
54. Drummond C, Holte R. C4. 5, class imbalance and cost sensitivity: why under-sampling beats over-sampling. In: Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II. 2003 Presented at: International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets II; Jul 21, 2003; Washington, DC, USA URL: https://www.site.uottawa.ca/~nat/Workshop2003/drummondc.pdf
55. Batista G, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor Newsl 2004 Jun 01;6(1):20-29. [doi: 10.1145/1007730.1007735]
56. 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]
57. Estabrooks A, Jo T, Japkowicz N. A multiple resampling method for learning from imbalanced data sets. Comput Intell 2004 Feb;20(1):18-36. [doi: 10.1111/j.0824-7935.2004.t01-1-00228.x]
58. Yuan X, Xie L, Abouelenien M. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recognit 2018 May;77:160-172. [doi: 10.1016/j.patcog.2017.12.017]
59. Wei Q, Shi B, Lo J, Carin L, Ren Y, Hou R. Anomaly detection for medical images based on a one-class classification. In: Proceedings of the Conference on Medical Imaging 2018: Computer-Aided Diagnosis. 2018 Presented at: Conference on Medical Imaging 2018: Computer-Aided Diagnosis; Feb 27, 2018; Houston, Texas, United States. [doi: 10.1117/12.2293408]
60. Ruff L, Vandermeulen R, Goernitz N, Deecke L, Siddiqui S, Binder A, et al. Deep one-class classification. In: Proceedings of the 35th International Conference on Machine Learning. 2018 Presented at: 35th International Conference on Machine Learning; Jul 10-15, 2018; Stockholm Sweden URL: http://proceedings.mlr.press/v80/ruff18a.html
61. Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th international joint conference on Artificial intelligence. 2001 Presented at: IJCAI'01: 17th international joint conference on Artificial intelligence; Aug 4, 2001; Seattle WA USA. [doi: 10.5555/1642194.1642224]
62. Sun Y, Kamel MS, Wong AK, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognit 2007 Dec;40(12):3358-3378. [doi: 10.1016/j.patcog.2007.04.009]
63. Maloof M. Learning when data sets are imbalanced and when costs are unequal and unknown. In: Proceedings of the ICML'2003 Workshop: Learning from Imbalanced Data Sets II. 2003 Presented at: ICML'2003 Workshop: Learning from Imbalanced Data Sets II; Aug 21, 2003; Washington, DC URL: https://www.site.uottawa.ca/~nat/Workshop2003/maloof-icml03-wids.pdf
64. Joshi M, Kumar V, Agarwal R. Evaluating boosting algorithms to classify rare classes: comparison and improvements. In: Proceedings of the 2001 IEEE International Conference on Data Mining. 2001 Presented at: IEEE International Conference on Data Mining; Nov 29-Dec 2, 2001; San Jose, CA, USA. [doi: 10.1109/icdm.2001.989527]
65. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag 2009 Jul;45(4):427-437. [doi: 10.1016/j.ipm.2009.03.002]
66. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning. 2006 Jun Presented at: ICML '06: 23rd international conference on Machine learning; Jun 25-29, 2006; Pittsburgh Pennsylvania USA. [doi: 10.1145/1143844.1143874]
67. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015 Mar 4;10(3):e0118432 [FREE Full text] [doi: 10.1371/journal.pone.0118432] [Medline: 25738806]
68. Drummond C, Holte R. What ROC Curves Can't Do (and Cost Curves Can). URL: https://www.site.uottawa.ca/~nat/Courses/csi5388/Presentations/cost_curves.pdf [accessed 2021-08-16]
69. Drummond C, Holte RC. Cost curves: an improved method for visualizing classifier performance. Mach Learn 2006 May 8;65(1):95-130. [doi: 10.1007/s10994-006-8199-5]
70. Baltruschat IM, Nickisch H, Grass M, Knopp T, Saalbach A. Comparison of deep learning approaches for multi-label chest x-ray classification. Sci Rep 2019 Apr 23;9(1):6381 [FREE Full text] [doi: 10.1038/s41598-019-42294-8] [Medline: 31011155]

71. Rajpurkar P, Joshi A, Pareek A, Chen P, Kiani A, Irvin J, et al. CheXpedition: investigating generalization challenges for translation of chest x-ray algorithms to the clinical setting. arXiv.org. 2020. URL: http://arxiv.org/abs/2002.11379 [accessed 2021-08-16]
72. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018 Nov 6;15(11):e1002683 [FREE Full text] [doi: 10.1371/journal.pmed.1002683] [Medline: 30399157]
73. Elgendi M, Nasir MU, Tang Q, Smith D, Grenier J, Batte C, et al. The effectiveness of image augmentation in deep learning networks for detecting COVID-19: a geometric transformation perspective. Front Med (Lausanne) 2021 Mar 1;8:629134 [FREE Full text] [doi: 10.3389/fmed.2021.629134] [Medline: 33732718]
74. Glocker B, Robinson R, Castro D, Dou Q, Konukoglu E. Machine learning with multi-site imaging data: an empirical study on the impact of scanner effects. arXiv.org. 2019. URL: http://arxiv.org/abs/1910.04597 [accessed 2021-08-16]
75. Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW. A theory of learning from different domains. Mach Learn 2009 Oct 23;79:151-175. [doi: 10.1007/s10994-009-5152-4]
76. Wang M, Deng W. Deep visual domain adaptation: a survey. Neurocomputing 2018 Oct 27;312:135-153. [doi: 10.1016/j.neucom.2018.05.083]
77. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. In: Domain Adaptation in Computer Vision Applications. Basel, Switzerland: Springer; Sep 13, 2017.
78. Long M, Zhu H, Wang J, Jordan M. Unsupervised domain adaptation with residual transfer networks. arXiv.org. 2017. URL: http://arxiv.org/abs/1602.04433 [accessed 2021-08-16]
79. Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial discriminative domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jul 21-26, 2017; Honolulu, HI, USA. [doi: 10.1109/cvpr.2017.316]
80. Dou Q, Castro DD, Kamnitsas K, Glocker B. Domain generalization via model-agnostic learning of semantic features. arXiv.org. 2019. URL: https://arxiv.org/abs/1910.13580 [accessed 2021-08-16]
81. Bousmalis K, Trigeorgis G, Silberman N, Krishnan D, Erhan D. Domain separation networks. arXiv.org. 2016. URL: http://arxiv.org/abs/1608.06019 [accessed 2021-08-16]
82. Li H, Pan S, Wang S, Kot A. Domain generalization with adversarial feature learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018 Presented at: IEEE/CVF Conference on Computer Vision and Pattern Recognition; Salt Lake City, UT, USA; Jun 18-23, 2018. [doi: 10.1109/cvpr.2018.00566]
83. Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation. In: Proceedings of the 30th International Conference on Machine Learning. 2013 Presented at: 30th International Conference on Machine Learning; Jun 16-21, 2013; Atlanta, Georgia URL: http://proceedings.mlr.press/v28/muandet13.html
84. Volpi R, Namkoong H, Sener O, Duchi J, Murino V, Savarese S. Generalizing to unseen domains via adversarial data augmentation. In: Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018). 2018 Presented at: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018); Dec 2-8, 2018; Montréal, Canada URL: https://papers.nips.cc/paper/2018/file/1d94108e907bb8311d8802b48fd54b4a-Paper.pdf
85. Peng X, Huang Z, Sun X, Saenko K. Domain agnostic learning with disentangled representations. arXiv.org. 2019. URL: http://arxiv.org/abs/1904.12347 [accessed 2021-08-16]
86. Žliobaitė I. Learning under concept drift: an overview. arXiv.org. 2010. URL: http://arxiv.org/abs/1010.4784 [accessed 2021-08-16]
87. Wang S, Minku LL, Yao X. A systematic study of online class imbalance learning with concept drift. IEEE Trans Neural Netw Learn Syst 2018 Oct;29(10):4802-4821. [doi: 10.1109/tnnls.2017.2771290]
88. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. ACM Comput Surv 2014 Apr;46(4):1-37. [doi: 10.1145/2523813]
89. Žliobaitė I, Pechenizkiy M, Gama J. An overview of concept drift applications. In: Big Data Analysis: New Algorithms for a New Society. New York City: Springer International; 2016.
90. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv.org. URL: http://arxiv.org/abs/1412.6572 [accessed 2021-08-16]
91. Moosavi-Dezfooli S, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Jun 27-30, 2016; Las Vegas, NV, USA. [doi: 10.1109/cvpr.2016.282]
92. Qayyum A, Qadir J, Bilal M, Al-Fuqaha A. Secure and robust machine learning for healthcare: a survey. IEEE Rev Biomed Eng 2020 Jul 31;14:156-180 [FREE Full text] [doi: 10.1109/rbme.2020.3013489]
93. Huang R, Xu B, Schuurmans D, Szepesvari C. Learning with a strong adversary. arXiv.org. 2016. URL: http://arxiv.org/abs/1511.03034 [accessed 2021-08-16]
94. Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv.org. 2015. URL: http://arxiv.org/abs/1412.5068 [accessed 2021-08-16]

95. Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. In: Proceedings of the Network and Distributed Systems Security Symposium (NDSS). 2018 Presented at: Network and Distributed Systems Security Symposium (NDSS); Feb 18-21, 2018; San Diego, CA. [doi: 10.14722/ndss.2018.23198]
96. Gao J, Wang B, Lin Z, Xu W, Qi Y. DeepCloak: masking deep neural network models for robustness against adversarial samples. arXiv.org. 2017. URL: http://arxiv.org/abs/1702.06763 [accessed 2021-08-16]
97. Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (SP). 2016 Presented at: 2016 IEEE Symposium on Security and Privacy (SP); May 22-26, 2016; San Jose, CA, USA. [doi: 10.1109/sp.2016.41]
98. Katz G, Barrett C, Dill D, Julian K, Kochenderfer M. Reluplex: an efficient SMT solver for verifying deep neural networks. In: Computer Aided Verification. Basel, Switzerland: Springer; 2017.
99. Ross A, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. arXiv.org. 2017. URL: http://arxiv.org/abs/1711.09404 [accessed 2021-08-16]
100. Bradshaw J, Matthews A, Ghahramani Z. Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks. arXiv.org. 2017. URL: http://arxiv.org/abs/1707.02476 [accessed 2021-08-16]
101. Nguyen L, Wang S, Sinha A. A learning and masking approach to secure learning. In: Decision and Game Theory for Security. Basel, Switzerland: Springer International; 2018.
102. Metzen J, Genewein T, Fischer V, Bischoff B. On detecting adversarial perturbations. arXiv.org. 2017. URL: http://arxiv.org/abs/1702.04267 [accessed 2021-08-16]
103. Lu J, Issaranon T, Forsyth D. SafetyNet: detecting and rejecting adversarial examples robustly. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice, Italy. [doi: 10.1109/iccv.2017.56]
104. Gopinath D, Katz G, Pasareanu C, Barrett C. DeepSafe: a data-driven approach for checking adversarial robustness in neural networks. arXiv.org. 2020. URL: http://arxiv.org/abs/1710.00486 [accessed 2021-08-16]
105. Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: attacks and defenses. arXiv.org. 2020. URL: http://arxiv.org/abs/1705.07204 [accessed 2021-08-16]
106. Song Y, Kim T, Nowozin S, Ermon S, Kushman N. PixelDefend: leveraging generative models to understand and defend against adversarial examples. arXiv.org. 2018. URL: http://arxiv.org/abs/1710.10766 [accessed 2021-08-16]
107. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science 2019 Mar 22;363(6433):1287-1289 [FREE Full text] [doi: 10.1126/science.aaw4399] [Medline: 30898923]
108. Chakraborty A, Alam M, Dey V, Chattopadhyay A, Mukhopadhyay D. Adversarial attacks and defences: a survey. arXiv.org. 2018. URL: http://arxiv.org/abs/1810.00069 [accessed 2021-08-16]
109. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 2018 Feb 19;6:14410-14430. [doi: 10.1109/access.2018.2807385]
110. Selvaraju R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2017 Presented at: 2017 IEEE International Conference on Computer Vision (ICCV); Oct 22-29, 2017; Venice, Italy. [doi: 10.1109/iccv.2017.74]
111. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016 Aug Presented at: KDD '16: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Aug 13-17, 2016; San Francisco California USA. [doi: 10.1145/2939672.2939778]
112. Buhrmester V, Münch D, Arens M. Analysis of explainers of black box deep neural networks for computer vision: a survey. arXiv.org. 2019. URL: http://arxiv.org/abs/1911.12116 [accessed 2021-08-16]
113. Campolo A, Sanfilippo M, Whittaker M, Crawford K. AI now 2017 report. AI Now. 2017. URL: https://ainowinstitute.org/AI_Now_2017_Report.pdf [accessed 2021-08-16]
114. Milli S, Schmidt L, Dragan A, Hardt M. Model reconstruction from model explanations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019 Presented at: FAT* '19: Conference on Fairness, Accountability, and Transparency; Jan 29-31, 2019; Atlanta GA USA. [doi: 10.1145/3287560.3287562]
115. Shokri R, Strobel M, Zick Y. On the privacy risks of model explanations. arXiv.org. 2021. URL: http://arxiv.org/abs/1907.00164 [accessed 2021-08-16]
116. Larrazabal AJ, Nieto N, Peterson V, Milone DH, Ferrante E. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proc Natl Acad Sci U S A 2020 Jun 09;117(23):12592-12594 [FREE Full text] [doi: 10.1073/pnas.1919012117] [Medline: 32457147]
117. Verma S, Rubin J. Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness. 2018 Presented at: FairWare '18: International Workshop on Software Fairness; May 29, 2018; Gothenburg Sweden. [doi: 10.1145/3194770.3194776]
118. Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv.org. 2018. URL: http://arxiv.org/abs/1808.00023 [accessed 2021-08-16]

119. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger K. On fairness and calibration. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017 Dec Presented at: NIPS'17: 31st International Conference on Neural Information Processing Systems; December 4 - 9, 2017; Long Beach California USA p. 5684-5693. [doi: 10.5555/3295222.3295319]

120. Benjamin R. Assessing risk, automating racism. Science 2019 Oct 25;366(6464):421-422. [doi: 10.1126/science.aaz3873] [Medline: 31649182]

121. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol 2019 Mar;20(3):405-410 [FREE Full text] [doi: 10.3348/kjr.2019.0025] [Medline: 30799571]

122. Mahajan V, Venugopal VK, Murugavel M, Mahajan H. The algorithmic audit: working with vendors to validate radiology-AI algorithms-how we do it. Acad Radiol 2020 Jan;27(1):132-135. [doi: 10.1016/j.acra.2019.09.009] [Medline: 31818381]

123. Collins GS, Reitsma JB, Altman DG, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br Med J 2015 Jan 07;350:g7594. [doi: 10.1136/bmj.g7594] [Medline: 25569120]

Abbreviations

AI: artificial intelligence
AUROC: area under the receiver operating characteristic curve
NLP: natural language processing

Edited by C Lovis; submitted 15.03.21; peer-reviewed by V Gupta, G Liu; comments to author 17.06.21; revised version received 29.06.21; accepted 10.07.21; published 09.09.21.

Please cite as:
Kulkarni V, Gawali M, Kharat A
Key Technology Considerations in Developing and Deploying Machine Learning Models in Clinical Radiology Practice
JMIR Med Inform 2021;9(9):e28776
URL: https://medinform.jmir.org/2021/9/e28776
doi:10.2196/28776
PMID:34499049

Viewpoint

Using a New Model of Electronic Health Record Training to Reduce Physician Burnout: A Plan for Action

Vishnu Mohan[1], MD, MBI; Cort Garrison[1], MD; Jeffrey A Gold[2], MD

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, United States
[2]Division of Pulmonary & Critical Care Medicine, Department of Medicine, Oregon Health & Science University, Portland, OR, United States

Corresponding Author:
Vishnu Mohan, MD, MBI
Department of Medical Informatics and Clinical Epidemiology
Oregon Health & Science University
3181 SW Sam Jackson Park Rd
Mail Code BICC
Portland, OR, 97239-3098
United States
Phone: 1 5034944469
Email: mohanv@ohsu.edu

Abstract

Physician burnout in the United States has been growing at an alarming rate, and health care organizations are beginning to invest significant resources in combating this phenomenon. Although the causes for burnout are multifactorial, a key issue that affects physicians is that they spend a significant proportion of their time interacting with their electronic health record (EHR) system, primarily because of the need to sift through increasing amounts of patient data, coupled with a significant documentation burden. This has led to physicians spending increasing amounts of time with the EHR outside working hours trying to catch up on paperwork ("pajama time"), which is a factor linked to burnout. In this paper, we propose an innovative model of EHR training using high-fidelity EHR simulations designed to facilitate efficient optimization of EHR use by clinicians and emphasize the importance of both lifelong learning and physician well-being.

(JMIR Med Inform 2021;9(9):e29374) doi:10.2196/29374

KEYWORDS

electronic health records; clinician burnout; EHR training; clinician wellness; after-hours EHR use; EHR; patient data; burnout; simulation; efficiency; optimization; well-being

Introduction

Physician burnout is a significant problem in the United States today. One study has suggested that over 50% of physicians have experienced at least one symptom of burnout; alarmingly, the authors also noted that the frequency of burnout increased by 10% in just three years (2011-2014) [1].

The risk of burnout has only intensified because of the additional stress on physicians caused by the COVID-19 pandemic [2], further underscoring the urgent need to find a way to mitigate this professional crisis.

Is There a Relationship Between Electronic Health Record Use and Burnout?

While the etiology is multifactorial, the electronic health record (EHR) has been strongly implicated in physician burnout [3], particularly because physicians spend a substantial proportion of their workday using the EHR. For example, primary care physicians spend more than half of their workday (nearly 6 hours) interacting with the EHR, both during and after regular patient care hours [4].

Commercial EHRs tend to be large, complex, clunky software that are often not praised for their ease of use. Needless to say, physician satisfaction with their EHR is generally low [5]. As many as 70% of EHR users have reported health information technology–related stress and a substantial proportion of physicians are unable to complete many of their EHR-related tasks when at work. Therefore, they end up spending an excessive amount of time [6] catching up on EHR "paperwork" (the irony of using this word in the context of the EHR is not lost on us) at home. The implications for this are significant: physicians who reported moderately high or excessive time spent on EHRs at home had almost twice the risk of burnout [7].

Why Do Physicians Spend So Much Time Catching Up on EHR Work After Hours?

The emphasis on clinical workflow efficiency (a phenomenon that has seen a sharp increase in attention after the advent of the EHR) coupled with the increasing complexity of the medical record has led to an exponential increase in the amount of patient data recorded in the EHR. Not only do primary care physicians spend half their working day at the computer, about half their time in the EHR is spent engaging in clerical and administrative tasks (eg, documentation, order entry, billing, and coding) and about a quarter of the remainder of their EHR time is spent managing their inbox [8]. The clerical burden associated with EHR use, a consequence of compliance and regulatory requirements, may play a key role in promoting physician burnout [9,10]. The amount of EHR time may increase with the inclusion of more genomic and consumer data into the patient record. Combine this with a rapid rise in the use of patient portals due to the COVID-19 pandemic and the result is a "perfect storm" of excessive data and cognitive burden [11]. Some of this can be mitigated by reducing administrative requirements using regulation directed from the federal level and optimizing clinical workflows. However, the continual increase in the information needs of physicians highlights the importance of ensuring that physicians are effectively trained in how to use the EHR to effectively and efficiently perform these tasks, thus minimizing pain points [12,13].

What Is the Root of the Problem?

One primary issue is that current models of EHR training are limited. As EHR use has become ubiquitous in health care, organizations have typically focused on providing initial training on EHR use to clinicians. These initial training offerings typically focus on basic EHR use but do not provide opportunities to gain workflow proficiency. One study has suggested that 43% of clinician users considered initial EHR training to be "less than adequate" and almost 95% felt that it could be improved [14].

Once they have completed basic EHR training, physicians then learn EHR skills on the job and typically improve their EHR use by the process of trial and error while interacting with the system interface, by gleaning nuggets of best practices from their peers, or by gaining additional insight when there is a significant system update that typically necessitates rolling out a new wave of EHR training. This is not only inefficient, but also offers no certainty that what is gleaned in this on-the-job fashion is truly the best way to use the system.

Additionally, current EHR training models are typically not tailored to a physician's unique workflow and information needs—in essence, the type of information, the way information is retrieved, and the specifics of documentation are different for, for example, an ambulatory obstetrician, a medical intensivist, and a trauma surgeon. The current model of EHR training, relying on a one-size-fits-all approach, is unable to accommodate for these variations in clinical workflows between specialties and locations.

What Interventions Have Been Proposed?

Areas for potential interventions include the following: (1) improving EHR-related training, (2) remodeling clinical workflows, and (3) redesigning the EHR to better reflect optimal workflows. In this paper, we present a viable model for transforming EHR education.

Some organizations have attempted to mitigate gaps in EHR training by organizing additional sessions to optimize clinician use of the EHR, either through refresher courses or retreats [15]. One organization has combined this with a structured, rapid assessment of workflow patterns and designed training that is informed by clinician feedback [16]. Others have used one-on-one or group proficiency training to improve self-reported EHR efficiency [10,17], while another organization described an individualized learning plan for physicians to improve their use of the EHR [18].

However, these solutions are, in essence, rescue therapies designed to "undo" the damage of poor initial training. They are also time intensive and for the most part focus on one specific domain, usually centered around improving efficiency in documentation or optimizing charge capture, and do not encompass the full spectrum of EHR activities encountered in a physician's specialty and workflow.

What Is Our Model of EHR Training?

Overview

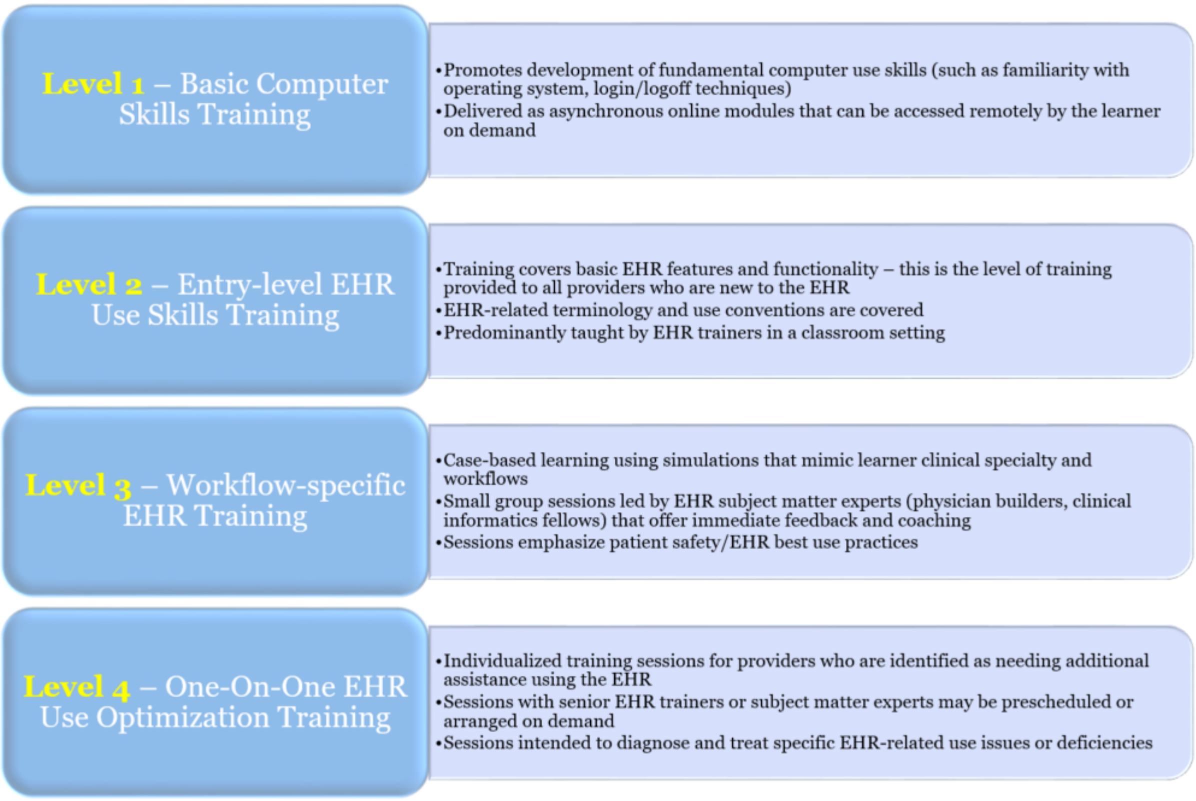
Over the past few years, we have conducted substantial research in the area of optimizing EHR use using simulations. We pioneered the use of high-fidelity simulation cases that replicate clinical cognitive loads and maintain EHR interface and documentation customizations created by the provider in their clinical environment. This allows learners to use the EHR just as they would when they deliver clinical care [19]. We then developed an intelligent simulation model to facilitate EHR training, which involved the use of an environment that replicated real-world EHR use [20] We also studied clinician interaction with the EHR by using eye tracking systems to assess EHR use during patient simulations [21].

Informed by our research, we have been able to show the utility of EHR-based simulation to improve efficiency and documentation in the patient record in a sustained manner [22] and identify and correct information gathering issues experienced by clinical end users [23]. Another key success factor is the importance of organizational investment in EHR training [24], particularly those that emphasize standards and personalization [25].

The results of our research coupled with those of others have allowed us to articulate a model of EHR training that allows efficient optimization of EHR use by clinicians while emphasizing the importance of both lifelong learning and physician well-being (Figure 1).

The model reorganizes EHR training into four levels, each capable of inducing a progressively higher degree of proficiency with respect to EHR use, and uses a stratified approach to the training process that compensates for prior EHR-related experience and proficiency (Figure 2). We believe that this model represents a paradigm shift in the EHR training universe, one that is more adaptive and responsive to clinician needs. This new model is currently being implemented at our institution.

Figure 1. Levels of electronic health record training. EHR: electronic health record.



Level 1 – Basic Computer Skills Training

- Promotes development of fundamental computer use skills (such as familiarity with operating system, login/logoff techniques)
- Delivered as asynchronous online modules that can be accessed remotely by the learner on demand

Level 2 – Entry-level EHR Use Skills Training

- Training covers basic EHR features and functionality – this is the level of training provided to all providers who are new to the EHR
- EHR-related terminology and use conventions are covered
- Predominantly taught by EHR trainers in a classroom setting

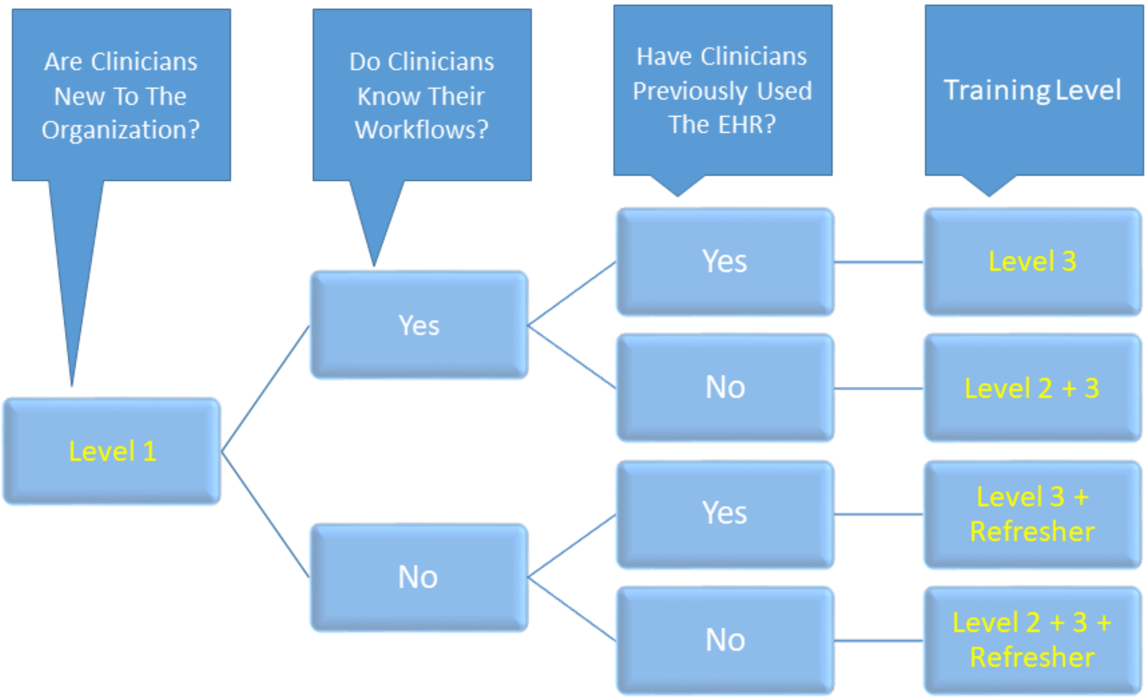
Level 3 – Workflow-specific EHR Training

- Case-based learning using simulations that mimic learner clinical specialty and workflows
- Small group sessions led by EHR subject matter experts (physician builders, clinical informatics fellows) that offer immediate feedback and coaching
- Sessions emphasize patient safety/EHR best use practices

Level 4 – One-On-One EHR Use Optimization Training

- Individualized training sessions for providers who are identified as needing additional assistance using the EHR
- Sessions with senior EHR trainers or subject matter experts may be prescheduled or arranged on demand
- Sessions intended to diagnose and treat specific EHR-related use issues or deficiencies

Figure 2. Algorithm for determining level of electronic health record training for physicians. EHR: electronic health record.


Are Clinicians New To The Organization?
Do Clinicians Know Their Workflows?
Have Clinicians Previously Used The EHR?
Training Level
Level 1
Yes
No
Yes
No
Yes
No
Level 3
Level 2 + 3
Level 3 + Refresher
Level 2 + 3 + Refresher

Level 1: Basic Computer Training

A surprising number of clinical end users struggle with the EHR because they lack basic computer skills—for example, a provider who exclusively uses Apple products as a consumer may have difficulty navigating the EHR because they are unfamiliar with the Windows operating system. These users may need to participate in learning that builds basic skills. This level of training can easily be delivered asynchronously, using online training modules that learners can complete on their own. Software-based solutions can also be used to correct some basic computer-related deficiencies (eg, improving the speed and accuracy of typing skills).

Level 2: General EHR Training

Second-level EHR training uses the typical one-to-many instructional model as commonly seen in EHR initial training sessions offered by most health care organizations. These are typically delivered in the classroom by EHR trainers and focus on explanations of basic features and functionality, as well as a demonstration of the fundamentals of EHR navigation, documentation, and order entry. Basic EHR training is an opportunity to emphasize standardized approaches to EHR use; training may also include highlighting high yield screens that clinicians can navigate to in order to optimize their EHR use [26]. Level 2 training is an appropriate entry point for all new users to an EHR system; however, the training should also allow a “test out” option for physicians who may be new to the organization but not to the EHR to avoid repetition of fundamentals.

Level 3: Workflow-Specific Training Using the EHR

Level 3 training integrates specialty-specific workflows and best practices related to clinical domains and patient safety. This is best achieved by using high-fidelity EHR simulations, where the clinical complexity of the environment can be duplicated in a replicable manner without endangering patient safety, as might occur if using real patient records in a production environment. Small groups of physicians complete simulation-based training sessions led by clinical informaticians, using specialty- and workflow-appropriate patient charts that have been imported into the simulation EHR environment prior to the activity. This model allows for instant debriefing as well as formative and summative feedback and coaching, and promotes retention of concepts learned during the session [22]. Coupling clinically relevant content to workflows familiar to the learner during EHR training is critical to successfully implementing this stage [27].

Level 4: One-on-one Training and Retraining

This level is characterized by tailored one-on-one EHR training and is typically reserved for providers who are identified as needing additional assistance. Level 4 sessions can be provided on demand or be scheduled to minimally disrupt the provision of clinical care by the learner. We use a simulated clinical case relevant to the provider’s clinical specialty to impart this level of EHR training, coupled with observation and eye tracking and keylogging software to differentiate information retrieval from cognitive issues that the provider may be encountering [23]. The session typically involves running through EHR use activities such as reviewing charts prior to rounding, documenting an encounter, and completing orders. One-on-one observations by the trainer (usually a clinical informatician or expert EHR user familiar with the clinical context) coupled with EHR use data recorded by the eye tracking and/or keylogging software allows the simulation team to effectively analyze EHR use, identify specific deficiencies, and prescribe a bespoke training solution to "diagnose and treat" EHR use issues.

Final Thoughts

The addition of an instructional designer to the training team greatly improves the quality of learning materials, particularly those that are offered asynchronously. We are obtaining continuing medical education and Maintenance of Certification credits for our EHR training, which promote compliance.

Importation of simulation cases into the training environment can be challenging, and the value of a team member who is trained in data importation into the EHR is critical to the success of any simulation-based training program.

COVID-19 has led to the virtualization of most nonclinical activities conducted by health care organizations, including EHR training. We believe that some forms of simulation-based training can only be provided in the face-to-face setting.

Finally, securing organizational commitment and allocation of adequate resources are often the most challenging elements in developing and deploying a comprehensive EHR training plan, but we believe these are also the most critical factors to ensure success.

Conflicts of Interest

None declared.

References

1. Shanafelt TD, Hasan O, Dyrbye LN, Sinsky C, Satele D, Sloan J, et al. Changes in Burnout and Satisfaction With Work-Life Balance in Physicians and the General US Working Population Between 2011 and 2014. Mayo Clin Proc 2015 Dec;90(12):1600-1613. [doi: 10.1016/j.mayocp.2015.08.023] [Medline: 26653297]
2. Amanullah S, Ramesh Shankar R. The Impact of COVID-19 on Physician Burnout Globally: A Review. Healthcare (Basel) 2020 Oct 22;8(4):421 [FREE Full text] [doi: 10.3390/healthcare8040421] [Medline: 33105757]
3. Downing NL, Bates DW, Longhurst CA. Physician Burnout in the Electronic Health Record Era: Are We Ignoring the Real Cause? Ann Intern Med 2018 May 08;169(1):50. [doi: 10.7326/m18-0139]
4. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Med 2017 Sep 11;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
5. Shanafelt TD, Dyrbye LN, Sinsky C, Hasan O, Satele D, Sloan J, et al. Relationship Between Clerical Burden and Characteristics of the Electronic Environment With Physician Burnout and Professional Satisfaction. Mayo Clin Proc 2016 Jul;91(7):836-848. [doi: 10.1016/j.mayocp.2016.05.007] [Medline: 27313121]
6. Kroth P, Morioka-Douglas N, Veres S, Pollock K, Babbott S, Poplau S, et al. The electronic elephant in the room: Physicians and the electronic health record. JAMIA Open 2018 Jul;1(1):49-56 [FREE Full text] [doi: 10.1093/jamiaopen/ooy016] [Medline: 31093606]
7. Gardner R, Cooper E, Haskell J, Harris D, Poplau S, Kroth P, et al. Physician stress and burnout: the impact of health information technology. J Am Med Inform Assoc 2019 Feb 01;26(2):106-114 [FREE Full text] [doi: 10.1093/jamia/ocy145] [Medline: 30517663]
8. Arndt BG, Beasley JW, Watkinson MD, Temte JL, Tuan W, Sinsky CA, et al. Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. Ann Fam Med 2017 Sep 11;15(5):419-426 [FREE Full text] [doi: 10.1370/afm.2121] [Medline: 28893811]
9. Holmgren AJ, Downing NL, Bates DW, Shanafelt TD, Milstein A, Sharp CD, et al. Assessment of Electronic Health Record Use Between US and Non-US Health Systems. JAMA Intern Med 2021 Feb 01;181(2):251-259. [doi: 10.1001/jamainternmed.2020.7071] [Medline: 33315048]
10. Downing NL, Bates DW, Longhurst CA. Physician Burnout in the Electronic Health Record Era. Ann Intern Med 2019 Feb 05;170(3):216-217. [doi: 10.7326/L18-0604] [Medline: 30716744]
11. Rand V, Coleman C, Park R, Karar A, Khairat S. Towards Understanding the Impact of EHR-Related Information Overload on Provider Cognition. Stud Health Technol Inform 2018;251:277-280. [Medline: 29968657]
12. Shannon D. Physician well-being: A powerful way to improve the patient experience. Physician Exec 2013;39(4):6-8, 10, 12. [Medline: 23923706]
13. Sinsky C, Colligan L, Li L, Prgomet M, Reynolds S, Goeders L, et al. Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties. Ann Intern Med 2016 Sep 06;165(11):753. [doi: 10.7326/m16-0961]
14. Rockswold PD, Finnell VW. Predictors of tool usage in the military health system's electronic health record, the Armed Forces Health Longitudinal Technology Application. Mil Med 2010 May;175(5):313-316. [doi: 10.7205/milmed-d-09-00286] [Medline: 20486501]

15. Dastagir M, Chin H, McNamara M, Poteraj K, Battaglini S, Alstot L. Advanced proficiency EHR training: effect on physicians' EHR efficiency, EHR satisfaction and job satisfaction. AMIA Annu Symp Proc 2012;2012:136-143 [FREE Full text] [Medline: 23304282]
16. DiAngi Y, Longhurst C, Payne T. Taming the EHR (Electronic Health Record) - There is Hope. J Fam Med 2016;3(6):1072 [FREE Full text] [Medline: 27830215]
17. Kirshner M, Salomon H, Chin H. An evaluation of one-on-one advanced proficiency training in clinicians' use of computer information systems. Int J Med Inform 2004 May;73(4):341-348. [doi: 10.1016/j.ijmedinf.2003.11.001] [Medline: 15135752]
18. Stevens L, DiAngi Y, Schremp J, Martorana M, Miller R, Lee T, et al. Designing An Individualized EHR Learning Plan For Providers. Appl Clin Inform 2017 Dec 20;08(03):924-935. [doi: 10.4338/aci-2017-04-0054]
19. Mohan V, Gold J. Collaborative intelligent case design model to facilitate simulated testing of clinical cognitive load. In: Workshop on Interactive Systems in Healthcare. 2014 Presented at: Workshop on Interactive Systems in Healthcare; 2014; Washington, DC.
20. Mohan V, Scholl G, Gold J. Intelligent Simulation Model To Facilitate EHR Training. AMIA Annu Symp Proc 2015;2015:925-932 [FREE Full text] [Medline: 26958229]
21. Gold JA, Stephenson LE, Gorsuch A, Parthasarathy K, Mohan V. Feasibility of utilizing a commercial eye tracker to assess electronic health record use during patient simulation. Health Informatics J 2016 Sep 26;22(3):744-757 [FREE Full text] [doi: 10.1177/1460458215590250] [Medline: 26142432]
22. March C, Scholl G, Dversdal R, Richards M, Wilson L, Mohan V, et al. Use of Electronic Health Record Simulation to Understand the Accuracy of Intern Progress Notes. J Grad Med Educ 2016 May;8(2):237-240 [FREE Full text] [doi: 10.4300/JGME-D-15-00201.1] [Medline: 27168894]
23. Mohan V, Scholl G, Gold JA. Use of EHR-based simulation to diagnose aetiology of information gathering issues in struggling learners: a proof of concept study. BMJ Simul Technol Enhanc Learn 2018 Apr 03;4(2):92-94 [FREE Full text] [doi: 10.1136/bmjstel-2017-000217] [Medline: 29657834]
24. Mohan V, Woodcock D, McGrath K, Scholl G, Pranaat R, Doberne JW, et al. Using Simulations to Improve Electronic Health Record Use, Clinician Training and Patient Safety: Recommendations From A Consensus Conference. AMIA Annu Symp Proc 2016;2016:904-913 [FREE Full text] [Medline: 28269887]
25. Longhurst C, Davis T, Maneker A, Eschenroeder H, Dunscombe R, Reynolds G, Arch Collaborative. Local Investment in Training Drives Electronic Health Record User Satisfaction. Appl Clin Inform 2019 Mar 15;10(2):331-335 [FREE Full text] [doi: 10.1055/s-0039-1688753] [Medline: 31091545]
26. Sakata KK, Stephenson LS, Mulanax A, Bierman J, Mcgrath K, Scholl G, et al. Professional and interprofessional differences in electronic health records use and recognition of safety issues in critically ill patients. J Interprof Care 2016 Sep 24;30(5):636-642 [FREE Full text] [doi: 10.1080/13561820.2016.1193479] [Medline: 27341177]
27. Miller ME, Scholl G, Corby S, Mohan V, Gold JA. The Impact of Electronic Health Record-Based Simulation During Intern Boot Camp: Interventional Study. JMIR Med Educ 2021 Mar 09;7(1):e25828 [FREE Full text] [doi: 10.2196/25828] [Medline: 33687339]

Abbreviations

EHR: electronic health record

Edited by C Lovis; submitted 04.04.21; peer-reviewed by C Longhurst; comments to author 05.06.21; revised version received 28.06.21; accepted 25.07.21; published 20.09.21.

Please cite as:
Mohan V, Garrison C, Gold JA
Using a New Model of Electronic Health Record Training to Reduce Physician Burnout: A Plan for Action
JMIR Med Inform 2021;9(9):e29374
URL: https://medinform.jmir.org/2021/9/e29374
doi:10.2196/29374
PMID:34325400

Original Paper

Sociotechnical Drivers and Barriers in the Consumer Adoption of Personal Health Records: Empirical Investigation

Umar Ruhi[1], BSc, MBA, PhD; Armin Majedi[2], BSc, MSc; Ritesh Chugh[3], GCTertEd, GDInfSys, MInfSys, PhD

[1]Business Analytics & Information Systems, Telfer School of Management, University of Ottawa, Ottawa, ON, Canada

[2]University of Ottawa, Ottawa, ON, Canada

[3]College of Information & Communication Technology, School of Engineering & Technology, Central Queensland University, Melbourne, Australia

Corresponding Author:
Umar Ruhi, BSc, MBA, PhD
Business Analytics & Information Systems
Telfer School of Management
University of Ottawa
55 Laurier East
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6135625800 ext 1990
Email: umar.ruhi@uottawa.ca

Abstract

Background: Increasingly popular in the health care domain, electronic personal health records (PHRs) have the potential to foster engagement toward improving health outcomes, achieving efficiencies in care, and reducing costs. Despite the touted benefits of PHRs, their uptake is lackluster, with low adoption rates.

Objective: This paper reports findings from an empirical investigation of the sociotechnical factors affecting the adoption of PHRs.

Methods: A research model comprising personal and technological determinants of PHR adoption was developed and validated in this study. Demographic, technographic, and psychographic data pertaining to the use of PHRs were collected through a web-based questionnaire for past, current, and potential users. Partial least squares-based structural equation modeling was used to estimate a structural model of cognitive and affective factors impacting intentions to use PHRs.

Results: The analysis revealed that in addition to the expected positive impact of a PHR system's usefulness and usability, system integration also positively affects consumers' intention to adopt. The results also suggest that higher levels of perceived usability and integration do not translate into higher levels of perceived usefulness. The study also highlights the importance of subjective norms, technology awareness, and technology anxiety as direct antecedents of the intention to adopt PHRs. The differential effects of the adoption factors are also discussed.

Conclusions: We hope that our study will contribute to the understanding of consumer adoption of PHRs and help improve the design and delivery of consumer-centric health care technologies. After discussing the implications for research, we provide suggestions and guidelines for PHR technology developers and constituents in the health care delivery chain.

(JMIR Med Inform 2021;9(9):e30322) doi:10.2196/30322

KEYWORDS

electronic personal health records; PHR; patient facing information systems; technology adoption; technology acceptance; consumer health informatics

Introduction

Background

Within the realm of health systems and applications, electronic personal health records (PHRs) represent a burgeoning technology that is gaining traction in many countries worldwide [1-5]. As a consumer-centric technology, a PHR can be defined as "an electronic application through which individuals can access, manage, and share their health information and that of others for whom they are authorized, in a private, secure, and confidential environment" [6]. In this regard, PHR systems comprise information and communication technologies that can potentially help all types of end users maintain health and wellness [7], and specifically facilitate patients to manage their ongoing illnesses [8].

In this paper, we characterize PHR technologies as those specifically pertaining to digitally stored health care information about an individual patient under the control of that patient or their caregiver [5,9]. This is in contrast to other technologies, such as electronic medical records (EMRs) and electronic health records (EHRs) that are typically maintained by health care providers or payor organizations [10]. Furthermore, our discussion applies to various forms of PHR systems identified in the extant literature, including stand-alone PHRs that require users to manually enter their health data and medical history [8,11,12], tethered PHRs that are offered as an extension of a health institution's back-end EHR or EMR [8,11,13], and interconnected PHRs that offer interoperability across various health information systems (HISs) [11,14].

Industry analysts have predicted great market potential for PHR-related technologies. For instance, according to studies conducted by the Markle Foundation, over 70% of US health care consumers believe that PHRs can improve health care quality [6,15]. Similarly, a study by Deloitte [16] highlighted that more than half of the US adult population may be interested in using web-based PHR services.

At the macro level, leveraging the potential value of PHRs in facilitating patient engagement and improving consumer health outcomes has been a key constituent of several government eHealth initiatives around the world. For example, in the United States, the Health Information Technology for Economic and Clinical Health Act established a meaningful use incentive program offering financial support to providers and health systems adopting EHR-related technologies [17]. Meaningful use stage 2 specifically calls for technologies that facilitate patient engagement in terms of personal health information management and care coordination, whereas stage 3 extends the requirements for these systems to include patient communication functions, patient education features, and interoperability with back-end EHRs [17,18]. Similarly, the European Union has funded several eHealth infrastructure projects with the aim of supporting personalized medicine, including the p-medicine EU project and the eHealthMonitor project [5]. Along similar lines in Canada, the Canada Health Infoway sponsors several federally funded projects to promote the adoption of consumer-focused digital health technologies ranging from health information records to patient-physician communication and remote patient monitoring [19].

Notwithstanding the industry forecasts about abundant consumer interest and government commitments to PHR technologies, the adoption of these technologies has been much slower than originally expected [4,20]. This disconnect between active interest and low actual use has been termed the *PHR paradox* [21]. Various reasons for lackluster adoption have been cited in the extant literature, often contradicting intuition, and sometimes with inconsistent findings across studies [22-25]. Consequently, many researchers have called for further studies in the area of consumer adoption of PHRs [21,22,25-27]. Our research aims to answer this call and further explore and clarify the role of sociotechnical factors in the adoption of PHRs.

In delineating the scope of investigation of this study, we would like to highlight our deliberate use of the term consumer instead of patient throughout the discussion. Our objective is to investigate factors that impact the adoption of PHRs from the perspective of all users who may be current as well as potential users of these systems. Toward this, we aim to include not only users who are currently receiving active care (patients), but also those who may simply be interested in maintaining their health information and medical history, or in using other nonclinical functionalities of PHRs (consumers). Other academic researchers and industry analysts have also commented on the distinction between patients and consumers, noting that consumers may include both current and prospective patients [28]. Moreover, consumers often have more decision-making flexibility than patients because the latter are primarily concerned with the management of their specific medical conditions [29-31].

By virtue of its orientation, this research study is principally situated in the field of consumer health informatics (CHI), a field concerned with health and health care-related preferences and information needs of consumers and associated medical and public health practitioners [32,33]. Technology applications such as PHRs, which can help empower consumers to manage their own health, constitute an important focus of attention in the CHI field [14,26,34]. In this study, we seek to explore various personal and technological factors that can affect the adoption of PHR tools and applications, identifying with the broad objectives for CHI research toward analyzing, modeling, and integrating consumer preferences into medical information systems (ISs) [35].

Related Work

Researchers who have investigated user adoption of PHRs have suggested that possible adoption barriers may be related to technology factors, such as privacy and security concerns, system usability, and poor integration with health care provider systems [36,37]. Furthermore, personal factors, such as inadequate technology competency, low technology awareness, unrealistic expectations, and presence of chronic medical conditions, have also been linked to the likelihood of adoption of these technologies [38-40]. Some of these factors have been empirically validated, but the results across investigations are often inconsistent [23-26,41-43].

Consequently, researchers have called for further empirical studies to explore and validate the role of specific PHR adoption factors. Multimedia Appendix 1 [8,22,24-26,39,41,44-57] provides a chronological summary of research studies in the area of PHR adoption and outlines key takeaways from each study. Specific calls for further research in each study are also highlighted.

Our review of the extant literature indicates that patients with chronic illnesses or disabilities, their caregivers, and people caring for older persons are more likely to adopt and use PHR technologies [15,44-46,58-61]. These groups of users will find PHR technologies useful as a communication tool to obtain personalized care from their clinicians [7,47-50,59] and as an organizational tool to help track patient health conditions, maintain medication lists, write patient diaries, and keep notes from physician consultations [7,8,41,49,50,60,62].

Current research also shows that factors such as computer anxiety, security and privacy concerns, and perceptions of usefulness are key determinants of PHR adoption across different consumer strata [22-24,43,51-54,63]. In contrast, research on several adoption factors, such as usability perceptions, consumer health literacy, and user self-efficacy, has shown varied and inconsistent results in the extant literature. For example, in multiple studies, Archer and Cocosila found different results pertaining to the impact of health-information seeking preferences and self-efficacy of individuals on the adoption of PHR systems [22,23,51].

In terms of key areas for further exploration, our review indicates the need for more research on PHR adoption along several lines. From the perspective of personal factors, there is a significant lack of empirical evidence on the role of social influence processes in PHR adoption. In our review, we found only two studies that investigated the role of subjective norms in the adoption of PHRs [52,64]. On the technology side, very few studies have empirically validated the role of usability perceptions and system integration attributes as part of the cognitive instrumental processes that impact PHR adoption. With respect to the former, only a few studies have investigated usability through the limited lens of perceived ease of use [24,52,55,56] despite anecdotal evidence and expert opinion that suggests that PHR usability includes additional dimensions [25,26,65]. Our study aims to address these gaps in the extant literature by conceptualizing these key factors and their relationships with other PHR adoption determinants. The next section describes our research model and its underlying constructs and hypotheses.

Research Model and Theoretical Underpinnings

Overview

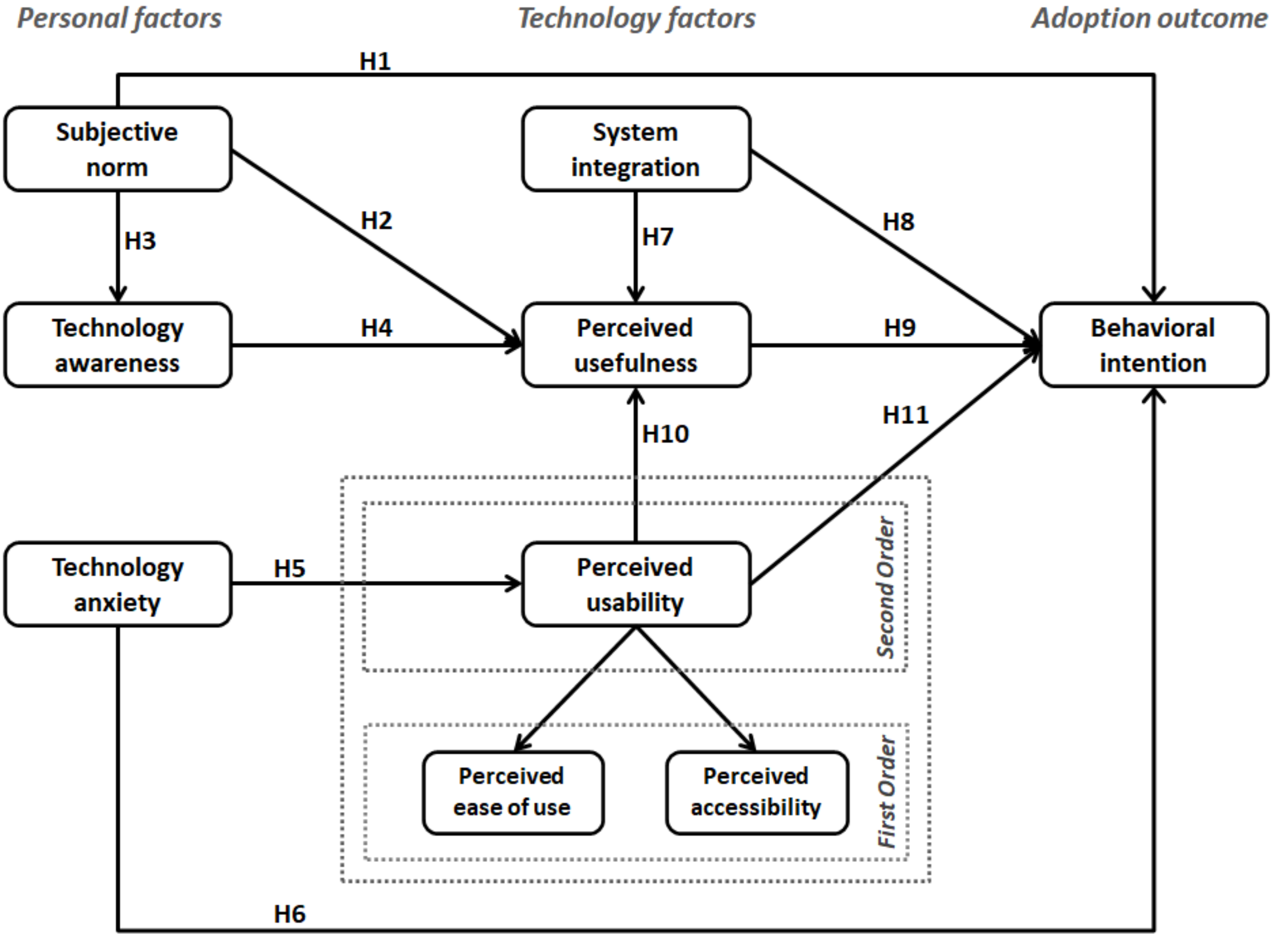
Notwithstanding the differences in results across some studies, researchers continue to investigate factors impacting consumer adoption of PHRs with the aim of improving our cumulative understanding of this phenomenon. As such, additional research in this area has been recommended by many researchers to further explore the impact of personal, technological, organizational, and environmental factors on consumer acceptance of PHR technologies, including patients and their caregivers [21,24,25,48,66].

This paper answers the call by theorizing and validating the role of various personal and technological factors as possible determinants of PHR adoption. We aim to contribute to the body of knowledge on the adoption of PHR systems by exploring sociotechnical factors that not only further clarify or complement those previously studied by other researchers, but also offer new avenues of inquiry. The scope of our investigation includes the study of subjective norms, technology awareness, and technology anxiety as personal factors affecting PHR adoption, and system integration, perceived usefulness, and perceived usability as technological antecedents of PHR adoption. These constructs and their definitions are provided in Table 1, and their posited interrelationships are shown in Figure 1. The theoretical justification for all research model constructs and hypotheses is outlined in the following subsections.

Table 1. Research model constructs.

Theme and constructs	Conceptual definition
Personal factors (determinants)	
Subjective norm	• The degree to which users perceive that most people who are important to them think they should or should not use the system [67,68]
Technology awareness	• An individual’s familiarity with the purpose and benefits of the technology [69,70]
Technology anxiety	• An individual’s apprehension or fear when confronted with the use of technology [71,72]
Technology factors (determinants)	
System integration	• Extent of connection and interoperability among technology components and sub-systems [73]
Perceived usefulness	• The degree to which users believes that using the system will help them toward achieving their desired goals [68,74]
Perceived usability (ease of use and accessibility)	• The degree of ease associated with the system [68,74] • Intuitive interface and information structure that is comprehensible and available when needed [75,76]
Adoption outcome (consequent)	
Behavioral intention	• The degree to which a person has formulated conscious plans to perform or not perform some specified future behavior [68]

Figure 1. Research model and construct definitions.



Subjective Norm

In technology adoption studies, the concept of subjective norm is appropriated to account for social influences that impact a potential user's decision to adopt and use a technology. The concept of subjective norm has its theoretical underpinnings in the theory of reasoned action, which defines it as "person's perception that most people who are important to him think he should or should not perform the behavior in question" [67]. In technology adoption studies, subjective norm represents perceived social pressure to use a new technology [68,71] and has been shown to be a significant determinant of behavioral intention to use a technology [77,78].

In the context of PHR system adoption, there is a dearth of research exploring the role of social influence on a user's decision to adopt these technologies. In our literature review, we identified one study that investigated subjective norms in the context of hardware-based (USB) PHRs within the specific regional context of Taiwan [52], and one study in Thailand, in which social influence was key in influencing the use of PHR [64]. As such, we expect subjective norms to play an even greater role as an antecedent of adoption for web-based PHRs, given that web-based technologies are likely to diffuse faster than hardware technologies. In addition, in this study, we aim to investigate whether subjective norm has only a direct impact on use intention, or whether it also plays an important role in internalizing the benefits of PHR technologies by affecting individual perceptions of the usefulness of these technologies. The following hypotheses related to subjective norms are posited in our research model:

- H1: Favorable subjective norms pertaining to the use of PHR technologies have a positive effect on the behavioral intention to use PHRs.
- H2: Favorable subjective norms pertaining to the use of PHR technologies have a positive effect on the perceived usefulness of PHRs.

Technology Awareness

Despite PHR technologies having been introduced more than a decade ago, research has found that there is a lack of awareness about them among many potential end users [49,54,56,79-82], thus inhibiting their use. This lack of awareness about PHR technologies has also been attributed to people having unrealistic expectations of these technologies [26,39,83], leading to their abandonment. A report by the Office of the National Coordinator for Health Information Technology [80] also found that people in the United States were especially unaware of stand-alone PHR offerings because they do not get similar promotional exposure as to health care institution-sponsored tethered PHR systems. Given the repercussions the lack of awareness can have on PHR adoption, several researchers have stated the need to address this research gap [26,79,84], and further posit calls for further research into the promotion of PHRs [82,85], including strategic wording [86] and educational or training programs [54,81]. Toward this, we draw upon the consideration of adoption studies conducted in the realm of other technologies to explore the role of technology awareness as a prerequisite to the development of perceptions about PHR usefulness [69,78].

In addition to exploring the role of technology awareness as a direct antecedent of perceived usefulness, we also explored its relationship with subjective norms. Research literature on the diffusion of innovation considers interpersonal relationships as an effective channel for creating awareness about an innovation [87,88]. These interpersonal channels can help create awareness by emphasizing the personal value of an innovation to a potential adopter [69]. We expect this to be the case for PHR technologies. The following two hypotheses related to technology awareness were tested in our research model:

- H3: Favorable subjective norms pertaining to the use of PHR technologies have a positive effect on technology awareness of PHRs.
- H4: Greater technology awareness of PHR technologies has a positive effect on the perceived usefulness of PHRs.

Technology Anxiety

Previous research in ISs shows technology anxiety to be a significant barrier to the adoption of new technologies [71,89], and the same findings have been echoed in research on the adoption of PHR systems [22,51]. However, a majority of PHR research to date simply considers the direct impact of technology anxiety on a user's intention to adopt PHRs without exploring its indirect effect on adoption through other key antecedents such as perceived ease of use. Past IS research shows technology anxiety to be an emotional anchor that leads to negative expectations of a technology [90], especially during the initial stages of its adoption. Previous IS studies have validated the importance of anxiety as an antecedent of perceived ease of use [91,92].

To address this gap in PHR adoption research, our model posits technology anxiety as an affective construct that affects the adoption of PHRs. We explored the direct link between anxiety and behavioral intention and its indirect effect on perceptions of usability (ease of use and accessibility). In doing so, our model attempts to capture the varying causes and effects of anxiety expressed in the extant literature on PHR adoption. These include inadequate technology literacy [27,59,81], individual uneasiness with setup of in-person authentication for tethered PHRs, lack of technical ability to integrate multiple data sources into stand-alone PHRs [80], or a general fear of technology [48,53]. In summary, we propose that technology anxiety potentially plays an important role in shaping cognitive responses toward PHR systems and directly affects behavioral intention to use these technologies. The following two hypotheses related to technology anxiety are proposed:

- H5: A higher level of technology anxiety has a negative effect on the perceived usability of PHRs.
- H6: A higher level of technology anxiety has a negative effect on the behavioral intention to use PHRs.

System Integration

Among the various contemporary PHR architectures, one may expect greater consumer interest in interconnected PHRs rather than stand-alone PHRs or even tethered PHRs. It is our position that with greater access to health and medical information available through multiple sources, consumers may be more motivated to use PHR systems. Such systems are likely to garner more interest through their *one-stop shopping* appeal, offering users a unified view of their health and medical information across the health care delivery chain.

Although many researchers and industry experts have commented on the lack of interoperability as a major barrier to consumer adoption [11,12,93-96], our literature review did not reveal any empirical substantiation of this conjecture. To address this issue, our research model incorporates system integration as a posited antecedent of perceived usefulness, as well as a direct determinant of behavioral intention. By exploring these relationships, we aim to investigate whether system integration aspects of PHRs are internalized through gradual system use, hence shaping user perceptions of usefulness, or whether the system integration factor is more prominent as an upfront reason to adopt or reject a PHR system. We offer the following two hypotheses related to system integration:

- H7: Greater system integration in PHR technologies has a positive impact on the perceived usefulness of PHRs.
- H8: Greater system integration in PHR technologies has a positive impact on the behavioral intention to use PHRs.

Perceived Usefulness

The extensive body of knowledge on the technology acceptance model (TAM) [74,97] shows that perceived usefulness is one of the strongest determinants of technology adoption [68,71,78]. Therefore, we expect perceived usefulness to be a strong determinant of PHR system adoption. Previous research on PHR adoption has validated the important role of perceived usefulness as a predictor of adoption [23,45,51,84,98]. In our model, we use perceived usefulness to signify performance expectancy in the use of PHRs, that is, the belief that using PHR will help in managing personal health. Furthermore, we also appropriately perceived usefulness as a cognitive response construct that is affected by other personal and technological determinants of PHR adoption. In addition to the previously posited hypotheses with perceived usefulness as the consequence (H2, H4, H7), we retained the conventional TAM hypothesis:

- H9: The higher perceived usefulness of PHR technologies has a positive impact on the behavioral intention to use PHRs.

Perceived Usability

Our final technological construct in the research model is theorized as a multidimensional factor consisting of the dimensions of perceived ease of use and perceived accessibility. The perceived usability construct in our model aims to capture the notion of effort expectancy associated with PHR systems, that is, the degree of ease associated with using PHRs.

The traditional view of the perceived ease of use construct in TAM also signifies effort expectancy [68,71]. However, research has shown that effort expectancy is usually a combination of ease of use and other contextual factors that shape end user perceptions about the relative difficulty of understanding and using the system [71]. In the context of PHR technologies, we believe that accessibility is a contextual factor that impinges effort expectancy. Research on PHR adoption factors indicates that aspects related to the intuitiveness of the user interface, understandability of information, availability through multiple channels (eg, desktop, web and mobile), and convenience of anytime anywhere access are important factors that affect individual perceptions of usability of PHR systems [8,65,99,100]. As such, our conceptualization of perceived accessibility attempts to assess the significance of these elements in determining end user perceptions of the usability of PHR technologies. To our knowledge, no previous research on PHR adoption has corroborated the role of accessibility in the acceptance of these technologies.

In conceptualizing perceived usability, we retain *ease of use* as an underlying dimension because it relates directly to other aspects of software usability, including end user efficiency and learnability with the system [101,102]. Furthermore, although previous research on PHRs has commented on the importance of ease of use for PHR adoption [53,56,83,95,96,103,104], very few studies have explored its role in the nomological network of other cognitive, affective, and behavioral factors [52,105]. On the basis of our multidimensional conceptualization of perceived usability, we propose the following two hypotheses:

- H10: Greater perceived usability of PHR technologies has a positive impact on the perceived usefulness of PHRs.
- H11: Greater perceived usability of PHR technologies has a positive impact on the behavioral intention to use PHRs.

Behavioral Intention

To characterize the adoption of PHRs, we used behavioral intention as the ultimate downstream construct in our research model. As a critical outcome of various cognitive and affective antecedents, this construct has its original basis within the theory of reasoned action [67], which conceptualizes it as a consequence of individual beliefs and as an antecedent of actual behavior. The construct has been commonly deployed in the IS literature to study the adoption of various types of technologies [74,97] including PHRs [22,23]. Furthermore, within the context of health behaviors, past research indicates that behavioral intention is significantly correlated with actual use [106-108]. Therefore, we expect greater behavioral intention to correspond to higher levels of actual use of PHR systems.

Overall, our research model aims to offer an inclusive basis for validating the role of three different types of determinants on PHR adoption—(1) individual differences, (2) system characteristics, and (3) social influence. Research models that include these categories of factors have been recommended as a practical foundation for investigating the adoption of new technologies [109]. It should be noted here that although we intend to be inclusive of these categories, we do not claim to be exhaustive over all possible adoption factors. As such, other adoption factors such as security and privacy concerns and health literacy have already been investigated in previous research studies, with largely consistent findings about the importance of these factors [22,51,84,110].

In terms of organization, our empirical methodology is described in terms of key procedures, and the results of our investigation are outlined. Finally, the discussion and conclusion sections offer an interpretation of the results, especially with respect to their implications for research and practice.

Methods

Survey Questionnaire Content

The research model posited in the previous section was validated through a quantitative empirical investigation using a web-based survey instrument. Details of the survey content, measurement scales, analysis procedures, and data collection techniques are presented below.

The survey comprised *demographic* information questions about the respondents' age, gender, and country of residence; *technographic* behavioral items related to respondents' experience and interest in using PHR technologies, as well as their preferences for different PHR features and functions, and *psychographic* questions pertaining to different constructs in the research model. For the latter, each construct in the research model was operationalized using multi-item psychometric scales with Likert-scale questions. Where possible, the items for a construct were adapted from previously validated measurement scales. We created new items for *system integration* and *perceived accessibility* constructs and modified the wording of items related to other constructs to align with the context of PHR systems.

To develop the two new scales, various qualitative and quantitative content validity assessment procedures were used, including concept elicitation interviews with subject matter experts (n=7) to generate representative and relevant measurement items; cognitive interviews with potential respondents from the target sampling frame (n=5) to ensure item relevance and clarity, and the final selection of measurement indicators based on item relevance ratings of subject matter experts, which were subsequently used to calculate item-level content validity indices (I-CVI). Drawing upon recommendations from the extant literature [111-113], a conservative cutoff value of 0.80 was used for item-level content validity indices to select items for the new scales. The 7 people in the subject matter expert panel included 2 faculty members from the health informatics domain at the authors' home institution, 1 health information technology business analyst working in a government agency, 2 doctoral students specializing in health information technology interoperability, one experienced end user of a PHR system, and a website manager of a patient portal of a health care institution.

At the end of the survey, participants were also invited to optionally respond to this open-ended question about PHR use: "Do you have any other comments about the use of personal health records (PHRs)? What factors do you consider to be important in your decision to start using or keep using technologies such as PHRs?"

The complete survey instrument was assessed for face validity through consultations with other HIS researchers, and construct validity for each theoretical construct was assessed through exploratory factor analysis of the pilot survey responses (n=20). Multimedia Appendix 2 [70,71] lists the final survey measurement items used for each construct in the research model.

Data Collection

Data for this study were collected through a web-based survey administered to actual and potential users of PHR technologies. Screening questions were asked at the beginning of the survey to determine different classes of respondents, and a brief overview of PHR technologies was offered to ensure qualified responses. As outlined in Multimedia Appendix 2, two alternative versions of questions were used to elicit responses from potential and actual (past or current) users of PHR systems.

The sampling techniques used were primarily based on convenience and self-selection. We recruited respondents who had basic familiarity with PHRs or similar tools for health care self-management. We used a two-pronged approach for data collection to ensure a cross section of potential PHR consumers. First, we solicited participation from current and past users of a PHR portal sponsored and supported by a teaching hospital (tethered PHR) in Ontario, Canada. In distributing our call for participation, we emphasized our interest in obtaining responses from current and past users of the PHR system. Second, calls for participation were also communicated through various web-based forums and social media groups dedicated to the discussion of health-related topics. To ensure a diverse selection of respondents, our sampling frame included both general health and wellness sites, as well as sites for chronic illness support groups. Once again, we underlined our goal of including responses from existing and potential users of PHR technologies.

Permission was sought from site administrators or forum moderators before posting our call for participation. In the case of the hospital PHR, our call for participation was distributed by the administrator to a mailing list of PHR users who had opted to receive news and information from the website at the time of their registration with the portal. No respondent incentives were offered for completing the survey.

The survey responses were collected over a 4-week period, with one reminder posted at each site with the original call for participation. Key suggestions from the Dillman tailored design method [114] were used to promote response rates for the survey. These included customizing the call for participation according to each site and posting personalized answers to any questions posted by potential respondents in a timely fashion. An interactive approach to collecting web-based survey data has been suggested by various researchers [115,116].

Because partial least squares (PLS) was the planned multivariate statistical analysis procedure in this study, the minimum sample size heuristic for PLS studies [117,118] was used for an a priori calculation of the required sample size. Using this heuristic, the minimum target sample size for this study was determined to be 60 valid responses. The heuristic suggests that the minimum sample size requirement for PLS- based models is determined by finding the larger of the following values: (1) 10 times the largest number of antecedent variables that affect any consequent in the model, or (2) 10 times the number of maximum indicators (manifest variables) in a latent variable in the model [117,118]. For the theoretical model under investigation, the *Behavioral Intention* construct has 5 direct antecedents, whereas *Perceived Usability* has the most indicators assigned to its measurement, specifically 6 items as shown in Multimedia Appendix 2. Therefore, the minimum target sample size for this study was determined to be 60 valid responses.

Analysis Procedures

Responses to demographic and technographic questions were analyzed using descriptive statistics and nonparametric statistical tests, and testing of research model constructs and hypotheses was conducted through exploratory factor analysis and PLS-based structural equation modeling (SEM) techniques. The PLS approach for SEM was selected for this study because of its suitability for small-sample exploratory research [119] and its flexibility with multivariate normality assumptions [120].

Testing for common method bias was achieved by using three different procedures—(1) the Harman post hoc one-factor test [121], (2) verification of latent variable correlations as recommended by Pavlou et al [122], and (3) the PLS-based common latent factor test suggested by Liang et al [123].

Results

Overview

A total of 224 responses were collected from various sources, including the hospital PHR portal, web-based forums, and social media groups in our sampling frame. After discarding partial responses, 168 responses were retained for further statistical analysis. This exceeded our minimum sample size target, as specified above. The results from our analysis of the survey responses are detailed in the following subsections.

Demographic and Technographic Highlights

Table 2 provides a summary of the basic demographic and technographic information from the survey responses analyzed. A significant proportion of respondents indicated familiarity with PHR technologies, with many respondents indicating current or past use of PHRs. Overall, 62% of the respondents self-identified themselves as either patients or caregivers.

Table 2. Key highlights from the respondent sample (n=168).

Demographic and technographic factors	Frequency, n (%)
Gender	
Female	96 (57.1)
Male	72 (42.9)
Age (years)	
18-25	22 (13.1)
26-35	31 (18.5)
36-45	66 (39.3)
46-55	28 (16.7)
55 or older	21 (12.5)
Respondents source	
PHR[a] portal	59 (35.1)
Online health communities	109 (64.9)
PHR familiarity and use	
Familiar	116 (69.1)
Current use	64 (38.1)
Past use	30 (17.9)
Health status identification	
Patients	66 (39.3)
Caregivers	39 (23.2)

[a] PHR: personal health record.

On the survey question pertaining to the importance of various health care issues, respondents consistently identified better clinical health care outcomes as the top priority for them. These were followed by issues surrounding better delivery of health care, including access and cost of health care, as well as better communication with physicians. Multimedia Appendix 3 shows the top 5 issues identified in our survey based on the mean importance of each health care issue. In addition, the figure shows the top 10 PHR features identified in our survey. On the basis of the mean utility scores ranging from 1 to 7, we can see that content-based features that allow consumers to exercise control over their medical information take precedence for most people, followed by connectivity features that facilitate patient-provider and patient-physician communication. Juxtaposed alongside each other, the health care issues that are top priority seem to be drivers for the use of many PHR features, for example, system features related to the management of chronic illnesses through tracking of health information and medical history were deemed extremely important overall.

The next section outlines the results of the assessment of psychographic variables in the posited research model. Following the two-step approach for SEM analysis suggested by Anderson et al [124], an examination of the measurement model was conducted before testing the structural model. Both the measurement and structural models were estimated using the SEM facilities of Smart PLS [125].

Measurement Model Assessment

The measurement model was assessed through a combination of exploratory factor analysis procedures and various tests for discriminant and convergent validities for the constructs in the research model.

We assessed our multidimensional operationalization of the *perceived usability* construct through exploratory factor analysis. Using principal axis factoring with promax rotation, a two-factor model emerged with 3 out of 7 items loading on the first factor and 3 on the second factor, all above the threshold of 0.70. One item that did not load well on either factor was dropped, and the scale was recalibrated with the remaining items, three corresponding to *perceived ease of use*, and three loadings on *perceived accessibility*. Subsequently, *perceived usability* was operationalized as a reflective higher-order factor structure in our model. To this end, we applied the repeated indicators (superblock) technique [126], which is the most commonly used approach for estimating hierarchical component models in PLS [127].

For our main measurement model, we inspected the loading and cross-loading of the indicators, as presented in Multimedia Appendix 4, Table S1. The highest loading for each measurement item (shown in bold) corresponds to its respective latent variable, and these loading values were higher in comparison to the item cross-loading on other model constructs. Moreover, except in one case, the substantive loading of each item on its construct exceeded the recommended threshold of

0.70, indicating item reliability [118]. In the case of item T_Anx_3, where the loading was slightly below the threshold, because the loading was rounded up to 0.70, the item was retained to ensure content validity. Overall, the assessment of loading and cross-loading demonstrated satisfactory reliability and discriminant validity at the item level.

We also followed the Fornell and Larcker guidelines [128] to ensure that the theoretical model constructs were all distinct. A visual inspection of Multimedia Appendix 4, Table S2 shows that for each construct, the square root of the average variance extracted (AVE; shown in bold on the diagonal) exceeds other interconstruct correlations. This demonstrates the discriminant validity of our measurement model at the construct level.

Various tests of convergent validity were performed through an assessment of quality indices, as shown in Multimedia Appendix 4, Table S3. As shown, the AVE value for each construct is higher than 0.5, indicating that at least 50% of the variance in each block of indicators can be attributed to the pertinent latent variables [118,128]. Moreover, the values of the Cronbach α are in the range of .60 or higher, thus demonstrating the internal reliability consistency of each construct [119]. Finally, the composite reliability values for each construct are higher than .70, which is the recommended cutoff to validate the internal reliability consistency of each construct relative to all other constructs in the model [128].

Finally, as part of the measurement model, we assessed the possibility of the common method bias using three different procedures.

First, the Harman post hoc one-factor test [121] was conducted. Principal component factor analysis (unrotated solution) revealed 6 factors extracted, with the first factor accounting for 27.3% of the variance. Common method bias was not deemed to be a serious problem with the data because multiple factors emerged, and no single factor accounted for a majority of the variance [121,129].

We subsequently applied the procedure specified by [122] and examined the latent variable correlation matrix from our PLS analysis. Usually, interconstruct correlations of over 0.90 indicate common method variance. In our data, the positive correlations ranged from 0.02 to 0.63, with no observed correlations exceeding the 0.90 threshold. Furthermore, the existence of several low correlations below 0.10 among some of the model constructs indicated that there was no single factor that influenced all constructs [122].

Finally, we used the PLS-based common method bias test suggested by Liang et al [123]. A method factor measured using indicators from all model constructs was added to the research model, and the variance of each item was then explained by its principal construct and method factor. Our results showed that the average variance explained by the principal constructs was 65.2%, whereas the average variance explained by the method factor was 21.5%. The ratio of substantive variance to method variance was approximately 3:1, suggesting that although there may be some common method variance, it does not account for the majority of the variance explained by the model.

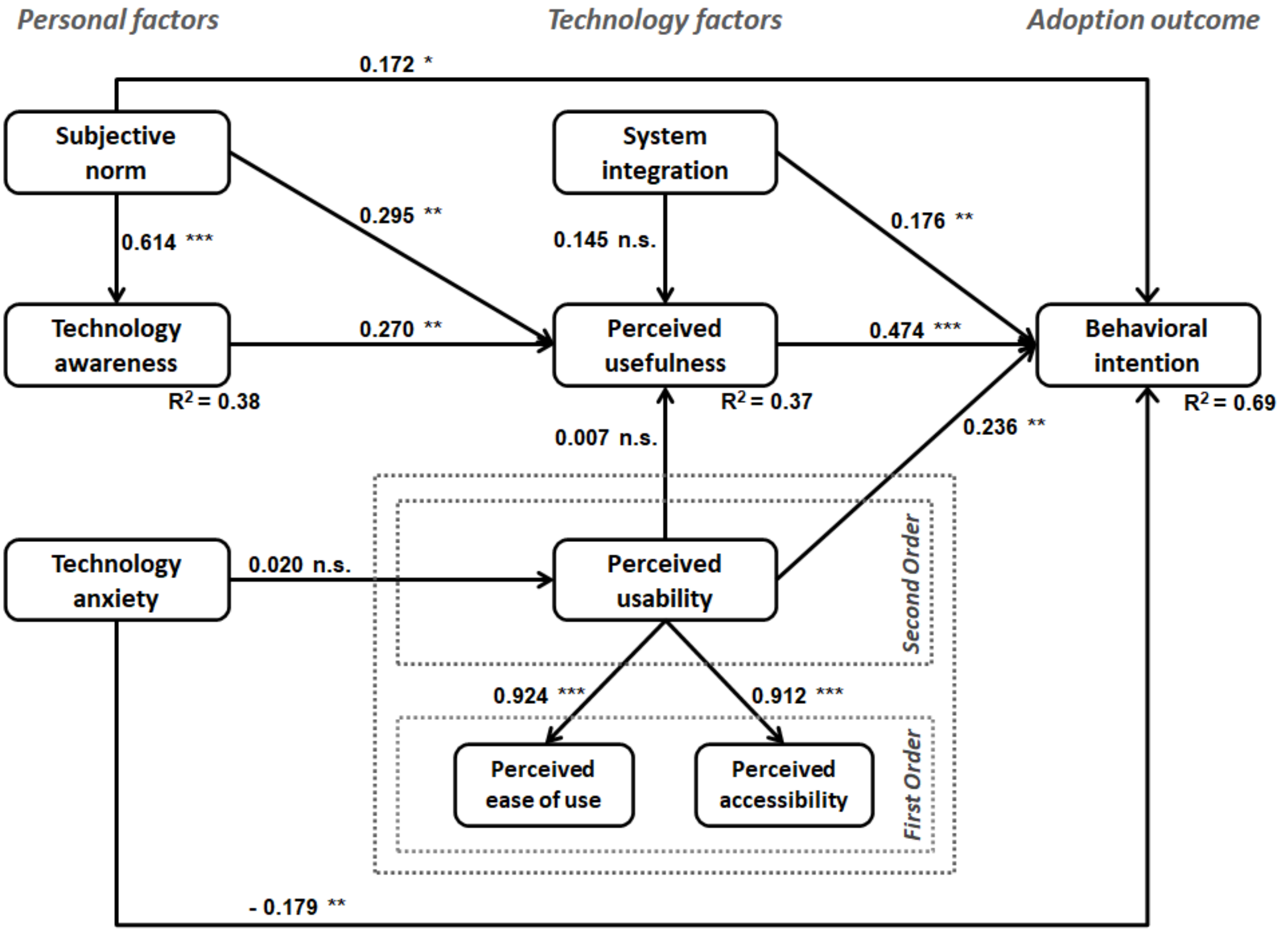
Overall, the assessment of the measurement model was deemed satisfactory in terms of item reliability and discriminant validity, and the model constructs were considered to be internally consistent as a measurement scale.

Structural Model Assessment

Following the measurement model assessment, the structural model was estimated to provide details of the strengths of the relationships among the latent constructs and the overall predictability of the endogenous latent variables in the model.

To estimate the structural model, path coefficients and significance levels were obtained by running PLS with bootstrapping using 1000 resamples. The structural model and *P* values are presented in Figure 2, with path β coefficients depicted along each path. As shown in Figure 2, 8 of the 11 hypotheses were supported with high degrees of confidence, and the model emerged as a good predictor of intention to adopt PHRs, as evidenced by the coefficient of determination (R^2) value of 0.69 for the ultimate criterion variable. The results are discussed next.

Figure 2. Estimated structural model. *P<.05, **P<.01, P<.001; n.s.: not significant at the .05 level.



With respect to *personal factors*, a significant relationship was validated between subjective norm and behavioral intention to adopt PHR technologies (H1 supported). In addition, as predicted, subjective norm had a significant positive effect on perceived usefulness and technology awareness (H2 and H3 supported). The path from technology awareness to perceived usefulness was also supported by the model (H4 supported). In terms of the effects of technology anxiety with PHRs, no significant association was found with perceived usability (H5 not supported), but a direct relationship with behavioral intention to use PHR technologies was validated (H6 supported).

The results pertaining to *technology factors* indicate that, contrary to expectations, system integration did not have a direct effect on the perceived usefulness of PHR technologies (H7 not supported). However, system integration was shown to have a direct impact on the user behavioral intention to adopt PHR technologies (H8 supported). As expected, perceived usefulness was shown to be a strong predictor of behavioral intention (H9 supported). With respect to perceived usability, we found an unexpected result of no significant relationship with perceived usefulness (H10 not supported). However, the direct effect of perceived usability on behavioral intention was validated in our model (H11 supported). Further comments on these results are provided in the *Discussion* section.

To determine the efficacy of the model in terms of predictability and goodness of fit (GoF), the coefficients of determination (R^2) and average communality (AVE) for each construct were evaluated. Together, these measures were used to calculate the global criterion of GoF, as recommended by several researchers [130,131]. Multimedia Appendix 4, Table S4 provide the R^2 values for all inner model constructs along with their average communalities and the calculated GoF index.

The R^2 values suggest that the model performed well for the endogenous variables pertaining to technology awareness, perceived usefulness, and behavioral intention. These coefficients of determination (R^2) explain the proportion of a construct's variance that can be predicted by antecedent constructs in the model. Most endogenous variables in the model compellingly exceed the minimum threshold of 0.10, indicating the usefulness of that variable in the model [132]. In terms of the ultimate criterion variable in the model, that is, behavioral intention to use PHRs, a significant portion of its variance (around 69%) can be explained by the posited research model.

To calculate the GoF index, the average communality of each construct is calculated as a weighted average of communality (AVE) based on the number of items in each construct taken as its weight [131]. Once calculated, the geometric mean of the average communality and the average R^2 can be calculated as specified in the GoF formula [131] in Multimedia Appendix 4, Table S4. The suggested baseline values for GoF are 0.1, 0.25, and 0.36 indicating small, medium, or large effect sizes, respectively [133]. As shown in Multimedia Appendix 4, Table S4, the GoF value of our model is 0.480, which exceeds the cutoff value of 0.36 for large effect sizes, allowing us to infer that the model performs well compared with the baseline values of effect sizes. Hence, it can be inferred that the structural model performed well overall.

On the basis of the evaluation of the measurement model validity and reliability, as well as the verification of predictive relevance and GoF of the structural model, we believe that the structural equation model was able to establish a strong basis for relationships posited in the research model hypotheses. Overall, the proposed model acts as an adequate predictor of behavioral intention to use PHRs.

Content Analysis of Open-ended Responses

As outlined earlier, we asked survey respondents to optionally provide comments about PHRs through textual responses to the question, “Do you have any other comments about the use of personal health records (PHRs)? What factors do you consider to be important in your decision to start using or keep using technologies such as PHRs?”

A total of 63 responses were submitted, and these were analyzed using simple content analysis techniques at the manifest level. In coding and classifying the qualitative data, we searched for themes or concepts related to the adoption of PHRs. An emergent coding technique was used whereby two researchers independently reviewed the responses and created a list of themes and codes. The list was consolidated after mutual consultation. Table 3 summarizes the comments that were classified using this procedure. In the table, we have only shown the three themes that are relevant to our research study—(1) consumer interest in PHR technology as a whole, (2) user interest in specific PHR features (grouped into categories), and (3) user concerns and potential barriers to adoption. It should be noted that each respondent could have contributed to multiple categories through their responses. Therefore, the frequency counts should be interpreted with caution.

Table 3. Content analysis summary for open-ended survey responses (n=63).

Themes and comments	Frequency, n (%)
General consumer interest in PHRs[a]	40 (64)
Support the idea of PHRs looking forward to their wider availability	16 (25)
PHRs are useful as they provide control or options to patients and their families	12 (19)
PHRs useful for chronic illness patients	14 (22)
Willing to pay or subscribe for PHR technologies	4 (6)
Interest in different PHR features (grouped into categories)	16 (25)
Medical information patient and provider records	8 (13)
Contact and communication with physician or provider	6 (10)
Decision support tools	4 (6)
Shared access and social networks	3 (5)
Concerns and barriers to adoption	18 (29)
Prefer data integration; unwilling to do manual data entry	12 (19)
Security and privacy concerns	11 (18)
Should be available through mobile apps	6 (10)

[a] PHR: personal health record**.**

On the whole, many respondents commented on the usefulness of PHR technologies as a whole and indicated their support and anticipation in adopting these technologies. Features related to the maintenance of medical information and online communication with physicians emerged as the most commonly cited PHR functions of interest. Interoperability, security, and privacy issues were frequently mentioned as key factors in the PHR adoption decision. Finally, some respondents stated their interest in using PHR technologies through mobile apps, hence alluding to the notion of accessibility as an important consideration for them.

Discussion

Overview

The results outlined in the previous section corroborate the general premise that a combination of personal and technological factors plays a role in determining the adoption of PHR technologies. In exploring these factors, our study has attempted to integrate constructs related to social influence beliefs (subjective norm), individual affective states (awareness and anxiety), and cognitive instrumental perceptions (system integration, perceived usability, and perceived usefulness) that potentially impact adoption behavior (behavioral intention) toward PHR technologies. This section provides an interpretation of the results and discusses the implications for research and practice.

Personal Factors

Our results indicate that a person's judgment of subjective norms pertaining to the use of PHR systems plays an important role in the adoption of these technologies through multiple cognitive and affective processes. Its direct impact on behavioral intention suggests that social influence plays a role in people's decision to adopt PHR technologies. The relatively weak association between subjective norms and behavioral intention can also be explained with reference to past research that shows that subjective norm does not factor prominently as a direct antecedent of behavioral intention in situations where the use of technology is voluntary [68]. This is certainly the case for most users of PHRs. In comparison, there is a stronger association between subjective norm and perceived usefulness, which suggests that internalization of social influence plays a far more important role in the context of PHR adoption. Internalization refers to the process by which a user incorporates the beliefs of an important referent into one's own belief structure [134]. What this means in the case of PHRs is that the consumers are more likely to develop their own perceptions about the usefulness of these technologies through information they receive from other important people, and this in turn can foster their intention to use PHR systems. Our model also shows that social influence through favorable subjective norms can improve an individual's awareness of PHR technologies. Overall, subjective norms seem to be an important factor in cognitive and affective mechanisms that allow an individual to make sense of the purpose and benefits of PHR systems.

The positive impact of technology awareness on perceived usefulness also alludes to a process of internalization whereby consumers' familiarity with the various use cases of PHR technologies allows them to develop beliefs about the technology's overall usefulness to them. Because the use of PHR systems is voluntary, it is reasonable to assume that consumers would take time to discover and understand the technology before deciding to adopt it. Once again, the relationship between subjective norms and technology awareness implies that observations and interactions with other people play an important role in this process.

Our results also support the critical role of technology anxiety as a determinant of PHR system adoption. Although no significant relationship emerged between technology anxiety and perceived usability, the construct exhibited a significant direct impact on behavioral intention to adopt PHR technologies. With respect to the former, although recent IS studies have shown anxiety to be an important antecedent of perceived ease of use [91,92], our study did not support this relationship. This finding can be attributed to a difference in the type of technology being investigated, as previous studies have generally focused on mandatory use or hedonic technologies. In the case of PHR applications, the technologies are expressly voluntary and instrumental for most consumers. It should also be noted that in adopting the current conceptualization of technology anxiety from the extant IS literature, we might have overlooked the multidimensional nature of anxiety as a psychological construct. Aligned with the IS literature, our construct conceptualization is reflective of anticipatory anxiety (apprehension preceding the use of PHR systems) rather than situational anxiety (distress during the use of PHR systems). The latter may indeed exhibit a relationship with perceived usability. Therefore, we recommend that the multidimensional nature of technology anxiety and its role in the adoption of PHR systems be investigated in future research.

Technology Factors

The construct of system integration was theorized in our research to measure the importance that users confer on interoperability (among PHRs and other back-end EHR or EMR systems) in their decisions to adopt PHR technologies. Our results demonstrate a positive association between consumer beliefs about PHR interoperability and the intention to adopt these technologies. However, the lack of support for the relationship between system integration features and perceptions of the usefulness of PHR technologies is counterintuitive. In the context of PHRs, it can be expected that better functionality of these systems in terms of connection and interoperability with other back-end systems would translate into better perceptions of the system's usefulness. This posture is supported by current research on PHR systems that consider a lack of integration between patient-facing systems and back-end eHealth systems as a barrier to adoption for both consumers and health care professionals [21,135].

These differential effects of system integration beliefs can be explained in the context of user expectations. It may be the case that given today's vast user experience with web-based tools and the pervasive deployment of web services linking different web-based systems, users simply expect PHR systems to be interoperable at the outset. Their common perception about PHRs would align with tethered and interconnected system models of PHRs, and it is these types of technologies that users are interested in adopting. Consumers may factor in these aspects of interoperability only during the initial stages of adoption, and these features are not internalized over time into higher-order cognitive states that represent perceptions of the usefulness of the system. As such, in our research model, the system integration construct is conceptualized in the form of initial expectations pertaining to PHR technologies, and it does not capture or measure aspects of assimilation of these technologies. Therefore, we suggest that future studies use a different approach to model the relationship between system integration and perceived usefulness. One possibility may be to draw upon the experience-disconfirmation theory, which has its roots in consumer behavior research [136], and posits that beliefs and behaviors result from the congruence between expectations and experiences [137].

Unlike many studies investigating technology adoption, our study did not find a significant relationship between perceived usability and perceived usefulness. Although this finding may be at odds with the general IS literature, the findings are not completely surprising in the specific context of PHR system adoption. Previous studies on PHR technology adoption have also shown varied results regarding the effects of perceived ease of use. Some studies confirm construct relationships as defined in the original TAM [107], whereas others contradict them [138]. We offer a possible explanation for this lack of a significant relationship by noting that PHR systems are characterized by their voluntary and instrumental use by potential end users, which requires an extended commitment on the part of end users to keep the system up-to-date and relevant and useful over time. Such systems have recently been the subject of IS research under the category of high maintenance ISs [139]. Initial research on high maintenance ISs contends that usability or ease of use may not be a prominent determinant of usefulness and behavioral intention, as its effect is usually superseded by the effect of other variables such as perceived maintenance effort [139]. In the case of PHR technologies, we expect a greater role for a construct, such as perceived maintenance effort, and future studies should incorporate this variable in their models.

In terms of direct effects on behavioral intention to adopt PHR systems, our results are consistent with the extant research literature. The role of perceived usability and perceived usefulness as antecedents of behavioral intention to adopt PHR systems was validated. Furthermore, having demonstrated internal reliability and construct validity, our integrated conceptualization of perceived usability as a combination of perceived ease and accessibility shows promise in the context of studying PHR technologies. Conceptualization lends support to many researchers' viewpoints on the synergistic relationship between usability and accessibility [2,65,99].

Responses to technographic questions and the open-ended questions in our survey also reveal consumer preferences for specific PHR features and functions. Our findings contribute to answering the call by other researchers, such as [57], who had asked future researchers to verify their own findings that consumers prefer health care process management support functions, such as communication and contact tools, more than other types of PHR tools. Our research verifies that these tools are among the most preferred tools, along with the category of tools that facilitate the maintenance of patient and provider records. Our findings show patient and provider records in PHRs to be the most preferred category of features, followed by communication and contact features. However, at least until the time when PHR adoption reaches its tipping point, we agree with other research studies that tools related to messaging, appointments, and prescription refills will remain the top-priority features for potential adopters of PHR technologies [140,141].

Implications for Research

Future studies should further investigate the role of norm internalization and technology assimilation as individual psychological processes affecting behavior toward PHR technologies. We suggest that the relationships among sociotechnical constructs reflect a gradual process in the development of beliefs about PHR technologies and their consequent adoption. For example, in this study, our results suggest that subjective norm and technology awareness are key constructs that affect the consolidation of individual and social values into higher-order cognitive beliefs about the purpose of the benefits of PHR technologies, that is, the internalization process. In the same vein, technology attributes, such as system integration and usability, feature more prominently in the affective and cognitive processes pertaining to technology assimilation. As a possible avenue for future investigations, we believe that incorporating mediating constructs from experience-disconfirmation theory could provide potentially valuable insights into PHR adoption research.

Future research should also seek to explore and validate the potentially multidimensional nature of some of the personal constructs posited in our theoretical model. Specifically, technology anxiety should be studied in terms of anticipatory and situational anxiety. We believe that both of these dimensions play an important role during the different stages of adoption of PHR technologies. Similarly, on the technology side, system integration should be operationalized through specific attributes of integration, such as single window patient information access, system-to-system health data sharing, and information communication capabilities, such as patient-physician exchanges. Doing so would also have the added benefit of deconstructing the specific needs and preferences of consumers in terms of their expectations of integration features and functions between PHR technologies and other HISs.

Our research also provides opportunities to improve health technology assessments. The conceptualization of the two new technology factors of system integration and perceived usability offered in our study may help enhance future systematic evaluations of health care technology. As highlighted earlier, our research shows that functionality, ease of use, and accessibility all play an important role in the adoption of PHR technologies.

Implications for Practice

In terms of practical implications, our research offers recommendations for PHR technology developers and designers, solution vendors, clinicians, and health policy makers.

Our study highlights the importance of system integration as a significant element affecting the initial decision to adopt PHRs. Technology developers should aim to incorporate interoperability as much as possible. Given the various challenges that exist in achieving seamless point-to-point integration across various types of HISs, developers and vendors should consider the use of health information exchanges as a viable alternative. Industry research suggests that health information exchanges may provide a practical solution to ensuring consumer access to comprehensive longitudinal health records from across the health care delivery chain [80,94].

PHR technology designers should also strive to incorporate accessibility as an element of overall PHR usability. In addition to being easy-to-learn and efficient-to-use, PHR tools should be available through a variety of channels, such as desktop, web, and mobile. Furthermore, PHR systems should facilitate help options and learning pathways to assist end user interactions with the technology features of PHR systems and to support a gradual learning curve. Technology should be developed in such a way as to mitigate anticipatory and situational anxiety with PHR technologies, and it should help end users feel in control of the system. A delineation of basic versus advanced features, context-sensitive suggestions for tasks and actions, and readily available technical support may help alleviate user anxiety and support the adoption of PHR systems [100].

Technology vendors can also help improve the uptake of their PHR systems by influencing personal affective and cognitive beliefs that influence behavior toward PHR technologies. For example, technology awareness can be improved and technology anxiety can be reduced by incorporating additional aspects of trialability and observability in PHR offerings. The availability of free trial versions or free subscriptions, interactive demonstration vignettes and how-to-use videos, access to a community of end users, and spotlights on positive consumer stories can provide useful mechanisms to help alleviate challenges pertaining to technology anxiety and awareness.

Health care providers and practitioners can help improve the uptake of PHR technologies by integrating these tools into clinical encounters and by engaging patients with the technology along various touchpoints in care delivery. The long-term benefits expected from the effective use of these technologies could potentially outweigh any increase in the short-term workload experienced by practitioners in helping promote these technologies to their patients.

From a policy perspective, relevant government agencies can prioritize training and development initiatives for people to become more proficient with the use of PHR systems. The target audience for such programs could include both consumers and health care professionals. The latter factor into the technology adoption process as key influencers as their engagement with patients and their endorsement of relevant PHR applications can accelerate the uptake of these technologies. Government-sponsored technology demonstrations can be administered at community centers or libraries to help improve literacy about PHR technologies, thereby improving consumer awareness, reducing anticipatory anxiety, and leading to greater adoption of these systems. Finally, at the infrastructure level, governments can accelerate the development of interoperability and health data interchange standards that would help make these systems more attractive to consumers and enable faster mainstream adoption.

Applicability Checks

To further confirm the relevance of our research to the health care sector, we performed applicability checks with several health care professionals, including two physicians, one hospital administrator, one system developer, and one health policy analyst. Applicability checks have been recommended as a useful method for researchers to improve communication between research and practice [142] and substantiate the practical relevance of research [143]. In conducting applicability checks for this research, we sought feedback on our research findings from health care professionals and asked them to comment on the importance of the issues identified in our research. A summary of key comments from the applicability check participants is included in Table 4. Overall, the participants indicated that research studies such as ours could potentially help improve the effective uptake of PHRs and produce efficiencies in the health care system. Furthermore, they commented on the potential of our research to help overcome PHR adoption barriers through actionable guidelines for the health care sector.

Table 4. Applicability checks and comments from health care professionals.

Health care professional	Perspective	Key comments
General practitioner (family medicine)	PHR[a] adoption for improved clinical health outcomes	• "PHRs can be great tools to allow patients to become more informed about their conditions and treatments." • "I believe that we can help patients get familiar with the benefits of PHRs and also help them get over their initial hesitation in using these tools."
Primary care physician (pediatrics)	PHR adoption for improved clinical health outcomes	• "I think PHR tools can be great for parents to keep track of their children's medical history. The information can later be handed over to children once they are able to manage it themselves." • "Once the technical hurdles are resolved, I think clinicians can play an important role in encouraging people to use these technologies. However, we [physicians] have to start using them too and lead by example."
Hospital administrator (director of operations)	PHR adoption for ensuring continuity of care	• "We currently provide access to patients to a limited part of their medical records. Having an integrated medical record across healthcare organizations can be very useful for timely interventions." • "As pointed out in this research, there are many technical obstacles to providing an integrated medical record and this probably hurts overall adoption."
Systems developer (EHR[b] systems; mobile health apps)	Functionality and usability requirements for PHR adoption	• "Providing access to patient information across organizations is a challenge. Various industry standards are attempting to resolve this issue. Once the problems are resolved, we can expect more user interest in these technologies." • "I agree that usability is more than just thinking about user-friendliness. Users today expect anytime anywhere access to information. This applies to PHRs as well."
Health policy analyst (digital health strategies)	eHealth initiatives and PHR adoption	• "There is a lot of work going on at the national and provincial levels to create the right conditions to support potential applications of PHR technologies." • "Suggestions made in this research can be useful in creating more awareness at the user level. Ultimately, we would like to see PHRs as a technology for all citizens."

[a] PHR: personal health record.

[b] EHR: electronic health record.

Study Limitations

As an exploratory study, our research has inherent limitations in terms of the posited research model. This includes hypotheses that did not emerge as significant. Another limitation of our study pertains to the use of convenience and self-selection sampling techniques. This may limit the generalizability of the results of this study. Furthermore, most of the respondents comprised a relatively younger age demographic from North America, and the results may not be representative of the general population.

We also note that by virtue of soliciting responses from a current PHR portal site, health information websites, and forums, our data were collected from respondents with some level of previous interest in health self-management. This limits our findings to current internet users with potentially higher health literacy and may not accurately account for the population of users with less exposure to health information or with less access to computing resources. Future research should include potential and actual users of PHR technologies through more diversified sources and utilize recruitment mechanisms to alleviate sampling bias.

Conclusions

Advancing the use of technologies in all walks of life is also increasing people's expectations of user-centered health care technologies. Consequently, consumer demand for PHR systems is likely to remain strong in the upcoming years. Recent academic and industry research on PHR systems has affirmed abundant consumer interest in these technologies [4,80,94].

The empirical research findings reported in this paper aim to contribute to the body of knowledge on consumer adoption of PHRs. To this end, we have attempted to explore and analyze possible factors contributing to what has been termed the *PHR paradox* [21], that is, despite their predicted benefits and considerable consumer interest, the adoption of PHRs has generally remained low. Our study also answers the call for researchers to investigate the facilitators and inhibitors of PHR adoption at multiple levels, including personal and technological [2,21,51,66].

By developing and validating a parsimonious research model comprising personal and technological determinants of PHR adoption, we were able to obtain several insights into the social influence and cognitive instrumental processes that impact consumer adoption of PHRs. Our results indicate that subjective norms, technology awareness, and technology anxiety are important factors that predict individual attitudes and beliefs about the usefulness of PHR systems and the ultimate adoption of these technologies. Our study also shows the differential effects of system integration capabilities and perceived usability on perceived usefulness and behavioral intention to adopt PHRs. Our characterization of PHR technologies in terms of their voluntary, instrumental, and high maintenance attributes has allowed us to make sense of some of the seemingly counterintuitive findings about technology antecedents of PHR adoption.

As such, our findings support the viewpoint of other researchers who contend that PHR technologies are complex innovations in which perceived attributes of technology are neither stable features nor sure determinants of adoption [21,95]. We encourage future research to examine the adoption of PHRs in a longitudinal fashion, exploring the role of different sociotechnical factors affecting users' cognitive and behavioral processes during the stages of internalization, assimilation, and maintenance of PHR systems.

We hope that the takeaways from our study will prove to be constructive in helping align PHR offerings more closely with consumer beliefs and attitudes, as well as their informational needs and functional requirements. This should help alleviate the risk of PHR technology rejection or abandonment.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Literature review summary.
[DOCX File , 44 KB - medinform_v9i9e30322_app1.docx]

Multimedia Appendix 2
Psychometric scales and measurement indicators.
[DOCX File , 26 KB - medinform_v9i9e30322_app2.docx]

Multimedia Appendix 3
Summary of responses to technographic questions.
[PNG File , 76 KB - medinform_v9i9e30322_app3.png]

Multimedia Appendix 4
Measurement and structural model assessment.
[DOCX File , 33 KB - medinform_v9i9e30322_app4.docx]

References

1. Pinciroli F, Pagliari C. Understanding the evolving role of the personal health record. Comput Biol Med 2015 Apr;59:160-163. [doi: 10.1016/j.compbiomed.2015.02.008] [Medline: 25726437]
2. Kahn JS, Aulakh V, Bosworth A. What it takes: characteristics of the ideal personal health record. Health Aff (Millwood) 2009;28(2):369-376. [doi: 10.1377/hlthaff.28.2.369] [Medline: 19275992]
3. Wolfson E. The personal health record. Healthline Networks Inc. URL: http://www.healthline.com/hlc/personal-health-record?micrositeId=30 [accessed 2021-05-15]
4. Ford EW, Hesse BW, Huerta TR. Personal health record use in the United States: forecasting future adoption levels. J Med Internet Res 2016 Mar 30;18(3):e73 [FREE Full text] [doi: 10.2196/jmir.4973] [Medline: 27030105]
5. Genitsaridi I, Kondylakis H, Koumakis L, Marias K, Tsiknakis M. Evaluation of personal health record systems through the lenses of EC research projects. Comput Biol Med 2015 Apr;59:175-185. [doi: 10.1016/j.compbiomed.2013.11.004] [Medline: 24315661]
6. Americans want benefits of personal health records. Markle. 2003. URL: https://www.markle.org/publications/950-americans-want-benefits-personal-health-records [accessed 2021-05-15]
7. Ruhi U, Chugh R. Utility, value, and benefits of contemporary personal health records: integrative review and conceptual synthesis. J Med Internet Res 2021 Apr 29;23(4):e26877 [FREE Full text] [doi: 10.2196/26877] [Medline: 33866308]
8. Tang PC, Ash JS, Bates DW, Overhage JM, Sands DZ. Personal health records: definitions, benefits, and strategies for overcoming barriers to adoption. J Am Med Inform Assoc 2006;13(2):121-126 [FREE Full text] [doi: 10.1197/jamia.M2025] [Medline: 16357345]
9. Kern J, Fister K, Polasek O. Active patient role in recording health data. In: Encyclopedia of Information Science and Technology, Second Edition. Hershey, Pennsylvania, United States: IGI Global; 2009.
10. Hoerbst A, Kohl CD, Knaup P, Ammenwerth E. Attitudes and behaviors related to the introduction of electronic health records among Austrian and German citizens. Int J Med Inform 2010 Feb;79(2):81-89. [doi: 10.1016/j.ijmedinf.2009.11.002] [Medline: 20031482]
11. Detmer D, Bloomrosen M, Raymond B, Tang P. Integrated personal health records: transformative tools for consumer-centric care. BMC Med Inform Decis Mak 2008 Oct 06;8:45 [FREE Full text] [doi: 10.1186/1472-6947-8-45] [Medline: 18837999]
12. Kharrazi H, Chisholm R, VanNasdale D, Thompson B. Mobile personal health records: an evaluation of features and functionality. Int J Med Inform 2012 Sep;81(9):579-593. [doi: 10.1016/j.ijmedinf.2012.04.007] [Medline: 22809779]

13. Anoshiravani A, Gaskin G, Kopetsky E, Sandborg C, Longhurst CA. Implementing an interoperable personal health record in pediatrics: lessons learned at an Academic Children's hospital. J Particip Med 2011 Jul 10;3:e30 [FREE Full text] [Medline: 21853160]
14. Johnson K, Jimison H, Mandl K. Consumer health informatics personal health records. In: Biomedical Informatics. London: Springer; 2013.
15. Americans overwhelmingly believe electronic personal health records could improve their health. Markle. 2008. URL: https://www.markle.org/sites/default/files/ResearchBrief-200806.pdf [accessed 2021-05-20]
16. The mobile personal health record: technology-enabled self-care. Deloitte. URL: https://www2.deloitte.com/us/2010mobilepersonalhealthrecord [accessed 2019-11-20]
17. Appari A, Johnson ME, Anthony DL. Meaningful use of electronic health record systems and process quality of care: evidence from a panel data analysis of U.S. acute-care hospitals. Health Serv Res 2013 Apr;48(2 Pt 1):354-375 [FREE Full text] [doi: 10.1111/j.1475-6773.2012.01448.x] [Medline: 22816527]
18. Goldzweig CL. Pushing the envelope of electronic patient portals to engage patients in their care. Ann Intern Med 2012 Oct 02;157(7):525-526 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00013] [Medline: 23027322]
19. Adler-Milstein J, Sarma N, Woskie LR, Jha AK. A comparison of how four countries use health IT to support care for people with chronic conditions. Health Aff (Millwood) 2014 Sep;33(9):1559-1566. [doi: 10.1377/hlthaff.2014.0424] [Medline: 25201660]
20. Gartrell K, Storr CL, Trinkoff AM, Wilson ML, Gurses AP. Electronic personal health record use among registered nurses. Nurs Outlook 2015;63(3):278-287 [FREE Full text] [doi: 10.1016/j.outlook.2014.11.013] [Medline: 25982768]
21. Nazi KM. The personal health record paradox: health care professionals' perspectives and the information ecology of personal health record systems in organizational and clinical settings. J Med Internet Res 2013 Apr 04;15(4):e70 [FREE Full text] [doi: 10.2196/jmir.2443] [Medline: 23557596]
22. Archer N, Cocosila M. Canadian patient perceptions of electronic personal health records: an empirical investigation. Commun Assoc Inf Syst 2014;34:A. [doi: 10.17705/1CAIS.03420]
23. Cocosila M, Archer N. Consumer perceptions of the adoption of electronic personal health records: an empirical investigation. In: Proceedings of the 18th Americas Conference on Information Systems. 2021 Presented at: 18th Americas Conference on Information Systems; Aug 9-11, 2012; Seattle, Washington, USA URL: https://aisel.aisnet.org/amcis2012/proceedings/ISHealthcare/10/
24. Emani S, Yamin CK, Peters E, Karson AS, Lipsitz SR, Wald JS, et al. Patient perceptions of a personal health record: a test of the diffusion of innovation model. J Med Internet Res 2012 Nov 05;14(6):e150 [FREE Full text] [doi: 10.2196/jmir.2278] [Medline: 23128775]
25. Pushpangadan S, Seckman C. Consumer perspective on personal health records: a review of the literature. Online J Nurs Informatics 2015 Jan;19(1):A [FREE Full text]
26. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. J Am Med Inform Assoc 2011;18(4):515-522 [FREE Full text] [doi: 10.1136/amiajnl-2011-000105] [Medline: 21672914]
27. Goldzweig CL, Orshansky G, Paige NM, Towfigh AA, Haggstrom DA, Miake-Lye I, et al. Electronic patient portals: evidence on health outcomes, satisfaction, efficiency, and attitudes: a systematic review. Ann Intern Med 2013 Nov 19;159(10):677-687. [doi: 10.7326/0003-4819-159-10-201311190-00006] [Medline: 24247673]
28. Folker G. The chronic need for connectivity: helping today’s aging heath care consumers help themselves. Consumer Health Informatics Summit, Ottawa, Canada. 2007. URL: https://web.archive.org/web/20180516112451/http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2007-Consumer-Health-Informatics-Summit.pdf [accessed 2021-04-12]
29. Scher D. Five differences between consumer and patient sensor technologies. Digital Health Corner. 2007. URL: https://davidleescher.wordpress.com/2014/06/02/five-differences-between-consumer-and-patient-sensor-technologies/ [accessed 2021-03-02]
30. Martineau M. The internet changes everything: lessons from other industries. Consumer Health Informatics Summit. 2007. URL: http://www.ehealthinformation.ca/wp-content/uploads/2014/07/2007-Consumer-Health-Informatics-Summit.pdf[accessed [accessed 2021-03-14]
31. Calabretta N. Consumer-driven, patient-centered health care in the age of electronic information. J Med Libr Assoc 2002 Jan;90(1):32-37 [FREE Full text] [Medline: 11838457]
32. Eysenbach G. Consumer health informatics. Br Med J 2000 Jun 24;320(7251):1713-1716 [FREE Full text] [Medline: 10864552]
33. Flaherty D, Hoffman-Goetz L, Arocha JF. What is consumer health informatics? A systematic review of published definitions. Inform Health Soc Care 2015 Mar;40(2):91-112. [doi: 10.3109/17538157.2014.907804] [Medline: 24801616]
34. Informatics: Research and practice. American Medical Informatics Association. URL: https://amia.org/about-amia/why-informatics/informatics-research-and-practice [accessed 2021-08-25]
35. Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the Internet age. J Med Internet Res 2001;3(2):E19 [FREE Full text] [doi: 10.2196/jmir.3.2.e19] [Medline: 11720961]
36. Gamble KH. Right on schedule. While there are no easy wins in IT, some say automated staff scheduling systems come close. Healthc Inform 2009 Sep;26(9):24, 26, 28-24, 26, 2-. [Medline: 19813570]

37. Maloney FL, Wright A. USB-based Personal Health Records: an analysis of features and functionality. Int J Med Inform 2010 Feb;79(2):97-111. [doi: 10.1016/j.ijmedinf.2009.11.005] [Medline: 20053582]
38. Wynia M, Dunn K. Dreams and nightmares: practical and ethical issues for patients and physicians using personal health records. J Law Med Ethics 2010;38(1):64-73. [doi: 10.1111/j.1748-720X.2010.00467.x] [Medline: 20446985]
39. Weitzman ER, Kaci L, Mandl KD. Acceptability of a personally controlled health record in a community-based setting: implications for policy and design. J Med Internet Res 2009 Apr 29;11(2):e14 [FREE Full text] [doi: 10.2196/jmir.1187] [Medline: 19403467]
40. Perzynski AT, Roach MJ, Shick S, Callahan B, Gunzler D, Cebul R, et al. Patient portals and broadband internet inequality. J Am Med Inform Assoc 2017 Sep 01;24(5):927-932 [FREE Full text] [doi: 10.1093/jamia/ocx020] [Medline: 28371853]
41. Ancker JS, Hafeez B, Kaushal R. Socioeconomic disparities in adoption of personal health records over time. Am J Manag Care 2016 Aug;22(8):539-540 [FREE Full text] [Medline: 27541700]
42. Fraccaro P, Vigo M, Balatsoukas P, Buchan IE, Peek N, van der Veer S. Patient portal adoption rates: a systematic literature review and meta-analysis. Stud Health Technol Inform 2017;245:79-83. [Medline: 29295056]
43. Javier SJ, Troszak LK, Shimada SL, McInnes DK, Ohl ME, Avoundjian T, et al. Racial and ethnic disparities in use of a personal health record by veterans living with HIV. J Am Med Inform Assoc 2019 Aug 01;26(8-9):696-702 [FREE Full text] [doi: 10.1093/jamia/ocz024] [Medline: 30924875]
44. King DK, Toobert DJ, Portz JD, Strycker LA, Doty A, Martin C, et al. What patients want: relevant health information technology for diabetes self-management. Health Technol 2012 Mar 5;2(3):147-157. [doi: 10.1007/s12553-012-0022-7]
45. Hsieh P, Lai H, Ku H, Ku W. Understanding middle-aged and elderly Taiwanese people’s acceptance of the personal health information system for self-health management. In: Human Aspects of IT for the Aged Population. Applications, Services and Contexts. Basel, Switzerland: Springer; 2017. [doi: 10.1007/978-3-319-58536-9_31]
46. Winkelman WJ, Leonard KJ, Rossos PG. Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness. J Am Med Inform Assoc 2005;12(3):306-314 [FREE Full text] [doi: 10.1197/jamia.M1712] [Medline: 15684128]
47. Smith AB, Odlum M, Sikka M, Bakken S, Kanter T. Patient perceptions of pre-implementation of Personal Health Records (PHRs): a qualitative study of people living with HIV in New York City. J HIV/AIDS Soc Serv 2012 Nov 19;11(4):406-423. [doi: 10.1080/15381501.2012.735166]
48. Luque AE, van Keken A, Winters P, Keefer MC, Sanders M, Fiscella K. Barriers and facilitators of online patient portals to personal health records among persons living with HIV: formative research. JMIR Res Protoc 2013 Jan 22;2(1):e8 [FREE Full text] [Medline: 23612564]
49. Clarke M, Karls K. Determining patient’s interest in patient portal use in a primary care clinic to improve portal adoption. In: Proceedings of the International Conference on Applied Human Factors and Ergonomics. 2019 Presented at: International Conference on Applied Human Factors and Ergonomics; Jul 24-28, 2019; Washington DC. [doi: 10.1007/978-3-030-20451-8_10]
50. Walker J, Ahern DK, Le LX, Delbanco T. Insights for internists: "I want the computer to know who I am". J Gen Intern Med 2009 Jun;24(6):727-732 [FREE Full text] [doi: 10.1007/s11606-009-0973-1] [Medline: 19412641]
51. Cocosila M, Archer N. Perceptions of chronically ill and healthy consumers about electronic personal health records: a comparative empirical investigation. BMJ Open 2014 Jul 23;4(7):e005304 [FREE Full text] [doi: 10.1136/bmjopen-2014-005304] [Medline: 25056975]
52. Jian W, Syed-Abdul S, Sood SP, Lee P, Hsu M, Ho C, et al. Factors influencing consumer adoption of USB-based Personal Health Records in Taiwan. BMC Health Serv Res 2012 Aug 27;12:277 [FREE Full text] [doi: 10.1186/1472-6963-12-277] [Medline: 22925029]
53. Kim E, Stolyar A, Lober WB, Herbaugh AL, Shinstrom SE, Zierler BK, et al. Challenges to using an electronic personal health record by a low-income elderly population. J Med Internet Res 2009 Oct 27;11(4):e44 [FREE Full text] [doi: 10.2196/jmir.1256] [Medline: 19861298]
54. Powell KR. Patient-perceived facilitators of and barriers to electronic portal use: a systematic review. Comput Inform Nurs 2017 Nov;35(11):565-573. [doi: 10.1097/CIN.0000000000000377] [Medline: 28723832]
55. Kahn JS, Hilton JF, Van Nunnery T, Leasure S, Bryant KM, Hare CB, et al. Personal health records in a public hospital: experience at the HIV/AIDS clinic at San Francisco General Hospital. J Am Med Inform Assoc 2010;17(2):224-228 [FREE Full text] [doi: 10.1136/jamia.2009.000315] [Medline: 20190069]
56. Gagnon M, Payne-Gagnon J, Breton E, Fortin J, Khoury L, Dolovich L, et al. Adoption of electronic personal health records in Canada: perceptions of stakeholders. Int J Health Policy Manag 2016 Jul 01;5(7):425-433 [FREE Full text] [doi: 10.15171/ijhpm.2016.36] [Medline: 27694670]
57. Agarwal R, Anderson C, Zarate J, Ward C. If we offer it, will they accept? Factors affecting patient use intentions of personal health records and secure messaging. J Med Internet Res 2013 Feb 26;15(2):e43 [FREE Full text] [doi: 10.2196/jmir.2243] [Medline: 23470453]
58. Lafky D, Horan T. Prospective personal health record use among different user groups: results of a multi-wave study. In: Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008). 2008 Presented at:

41st Annual Hawaii International Conference on System Sciences (HICSS 2008); Jan 7-10, 2008; Waikoloa, HI, USA. [doi: 10.1109/hicss.2008.363]

59. Huvila I, Enwald H, Eriksson-Backa K, Hirvonen N, Nguyen H, Scandurra I. Anticipating ageing: older adults reading their medical records. Inf Proc Manag 2018 May;54(3):394-407. [doi: 10.1016/j.ipm.2018.01.007]
60. Torrens E, Walker SM. Demographic characteristics of Australian health consumers who were early registrants for opt-in personally controlled electronic health records. Health Inf Manag 2017 Sep;46(3):127-133. [doi: 10.1177/1833358317699341] [Medline: 28537210]
61. Fernandez N, Copenhaver DJ, Vawdrey DK, Kotchoubey H, Stockwell MS. Smartphone use among postpartum women and implications for personal health record utilization. Clin Pediatr (Phila) 2017 Apr;56(4):376-381. [doi: 10.1177/0009922816673438] [Medline: 27798390]
62. Miles RC, Hippe DS, Elmore JG, Wang CL, Payne TH, Lee CI. Patient access to online radiology reports: frequency and sociodemographic characteristics associated with use. Acad Radiol 2016 Sep;23(9):1162-1169. [doi: 10.1016/j.acra.2016.05.005] [Medline: 27287715]
63. Dhanireddy S, Walker J, Reisch L, Oster N, Delbanco T, Elmore JG. The urban underserved: attitudes towards gaining full access to electronic medical records. Health Expect 2014 Oct;17(5):724-732 [FREE Full text] [doi: 10.1111/j.1369-7625.2012.00799.x] [Medline: 22738155]
64. Lalitaphanit K, Theeraroungchaisri T. Factors affecting pharmacy customers' decision to use personal health records via smartphone. Thai J Pharm Sci (Supplement) 2016;40:163-167 [FREE Full text]
65. Siek K, Khan D, Ross S. A usability inspection of medication management in three personal health applications. In: Proceedings of First International Conference on Human Centered Design HCD. 2009 Presented at: First International Conference on Human Centered Design HCD; Jul 19-24, 2009; San Diego, CA, USA. [doi: 10.1007/978-3-642-02806-9_16]
66. Logue MD, Effken JA. Modeling factors that influence personal health records adoption. Comput Inform Nurs 2012 Jul;30(7):354-362. [doi: 10.1097/NXN.0b013e3182510717] [Medline: 22525046]
67. Fishbein M, Ajzen I. Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research. Boston, MA: Addison-Wesley; 1975.
68. Venkatesh V, Davis FD. A theoretical extension of the technology acceptance model: four longitudinal field studies. Manag Sci 2000 Feb;46(2):186-204. [doi: 10.1287/mnsc.46.2.186.11926]
69. Agarwal R, Prasad J. The antecedents and consequents of user perceptions in information technology adoption. Decis Support Syst 1998 Jan;22(1):15-29. [doi: 10.1016/s0167-9236(97)00006-7]
70. Chan F, Thong J, Venkatesh V, Brown S, Hu P, Tam K. Modeling citizen satisfaction with mandatory adoption of an E-Government technology. J Assoc Inf Syst 2010 Oct;11(10):519-549. [doi: 10.17705/1jais.00239]
71. Venkatesh V, Morris M, Davis G, Davis F. User acceptance of information technology: toward a unified view. MIS Q 2003 Sep;27(3):425-478. [doi: 10.2307/30036540]
72. Simonson MR, Maurer M, Montag-Torardi M, Whitaker M. Development of a standardized test of computer literacy and a computer anxiety index. J Educ Comput Res 1987 May 01;3(2):231-247. [doi: 10.2190/7chy-5cm0-4d00-6jcg]
73. Iacovou CL, Benbasat I, Dexter AS. Electronic data interchange and small organizations: adoption and impact of technology. MIS Q 1995 Dec;19(4):465-485. [doi: 10.2307/249629]
74. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 1989 Sep;13(3):319-340. [doi: 10.2307/249008]
75. Teo H, Chan H, Wei K, Zhang Z. Evaluating information accessibility and community adaptivity features for sustaining virtual learning communities. Int J Hum Comput Stud 2003 Nov;59(5):671-697. [doi: 10.1016/s1071-5819(03)00087-9]
76. Rice RE, Shook DE. Access to, usage of, and outcomes from an electronic messaging system. ACM Trans Inf Syst 1988 Jul;6(3):255-276. [doi: 10.1145/45945.214325]
77. Mathieson K. Predicting user intentions: comparing the technology acceptance model with the theory of planned behavior. Inf Syst Res 1991 Sep;2(3):173-191. [doi: 10.1287/isre.2.3.173]
78. Jeyaraj A, Rottman JW, Lacity MC. A review of the predictors, linkages, and biases in IT innovation adoption research. J Inf Technol 2006 Feb 01;21(1):1-23. [doi: 10.1057/palgrave.jit.2000056]
79. Xie J, Rooholamini S, Pearson D, Bergman D, Winograd T. Barriers and facilitators of personal health record adoption: A study of low-income families with children with special health care needs. Stanford University. 2014. URL: https://forum.stanford.edu/events/posterslides/BarriersandFacilitatorsofPersonalHealthRecordAdoptioninLowIncomeFamilieswithChildrenwithSpecialHealthCareNeeds.pdf [accessed 2021-02-09]
80. PHR Ignite - Action: Healthinsight final report assessing the current environment and functionalities of PHR systems. Office of the National Coordinator for Health Information Technology. 2014. URL: https://www.healthit.gov/sites/default/files/healthinsight_phr_ignite_finalreport.pdf [accessed 2021-03-15]
81. Nahm E, Zhu S, Bellantoni M, Keldsen L, Charters K, Russomanno V, et al. Patient portal use among older adults: what is really happening nationwide? J Appl Gerontol 2020 Apr;39(4):442-450 [FREE Full text] [doi: 10.1177/0733464818776125] [Medline: 29779422]

82. Mishuris RG, Stewart M, Fix GM, Marcello T, McInnes DK, Hogan TP, et al. Barriers to patient portal access among veterans receiving home-based primary care: a qualitative study. Health Expect 2015 Dec;18(6):2296-2305 [FREE Full text] [doi: 10.1111/hex.12199] [Medline: 24816246]
83. Kunene K, Zysk K. Healthcare consumers' voluntary adoption and non-adoption of electronic personal health records. In: Proceedings of the 27th Australasian Conference on Information Systems. 2016 Presented at: 27th Australasian Conference on Information Systems; 2016; University of Wollongong Faculty of Business URL: https://aisel.aisnet.org/acis2016/78/
84. Whetstone M, Goldsmith R. Factors influencing intention to use personal health records. Intl J Pharm Health Mark 2009 Apr 03;3(1):8-25. [doi: 10.1108/17506120910948485]
85. Kim H, Chang CF. Effectiveness of using personal health records to improve recommended breast cancer screening and reduce racial and geographic disparities among women. J Cancer Educ 2020 Jul 09:A. [doi: 10.1007/s13187-020-01821-2] [Medline: 32648239]
86. Glowacki EM. Prompting participation in health: fostering favorable attitudes toward personal health records through message design. Patient Educ Couns 2016 Mar;99(3):470-479. [doi: 10.1016/j.pec.2015.10.004] [Medline: 26531806]
87. Rogers E. Diffusion of Innovations, 3rd Edition. Mumbai: Free Press; 1983.
88. Rogers E. Diffusion of Innovations, 4th Edition. Mumbai: Free Press; Feb 1, 1995.
89. Igbaria M, Schiffman SJ, Wieckowski TJ. The respective roles of perceived usefulness and perceived fun in the acceptance of microcomputer technology. Behav Inf Technol 1994 Nov;13(6):349-361. [doi: 10.1080/01449299408914616]
90. Venkatesh V. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. Inf Syst Res 2000 Dec;11(4):342-365. [doi: 10.1287/isre.11.4.342.11872]
91. George Saadé R, Kira D. Computer anxiety in e-learning: the effect of computer self-efficacy. J Inf Technol Educ Res 2009 Jan;8:177-191. [doi: 10.28945/166]
92. Hackbarth G, Grover V, Yi MY. Computer playfulness and anxiety: positive and negative mediators of the system experience effect on perceived ease of use. Inf Manag 2003 Jan;40(3):221-232. [doi: 10.1016/s0378-7206(02)00006-x]
93. Studeny J, Coustasse A. Personal health records: is rapid adoption hindering interoperability? Perspect Health Inf Manag 2014 Jul 1;11:1e [FREE Full text] [Medline: 25214822]
94. Key considerations for HIE-based personal health records. Venesco. 2015. URL: https://www.healthit.gov/sites/default/files/phrkeyconsiderations.pdf [accessed 2021-02-06]
95. Greenhalgh T, Hinder S, Stramer K, Bratan T, Russell J. Adoption, non-adoption, and abandonment of a personal electronic health record: case study of HealthSpace. Br Med J 2010 Nov 16;341:c5814 [FREE Full text] [doi: 10.1136/bmj.c5814] [Medline: 21081595]
96. Personal health records to improve health information exchange and patient safety. In: Advances in Patient Safety: New Directions and Alternative Approaches. US: Agency for Healthcare Research and Quality; 2008.
97. Davis FD, Bagozzi RP, Warshaw PR. User acceptance of computer technology: a comparison of two theoretical models. Manag Sci 1989 Aug;35(8):982-1003. [doi: 10.1287/mnsc.35.8.982]
98. Sääskilahti M, Aarnio E, Lämsä E, Ahonen R, Timonen J. Use and non-use of a nationwide patient portal – a survey among pharmacy customers. J Pharm Heal Serv Res 2020 Nov;11(4):335-342. [doi: 10.1111/jphs.12368]
99. Goldberg L, Lide B, Lowry S, Massett HA, O'Connell T, Preece J, et al. Usability and accessibility in consumer health informatics current trends and future challenges. Am J Prev Med 2011 May;40(5 Suppl 2):187-197. [doi: 10.1016/j.amepre.2011.01.009] [Medline: 21521594]
100. Lee M, Delaney C, Moorhead S. Building a personal health record from nursing perspective. Stud Health Technol Inform 2006;122:25-29. [Medline: 17102211]
101. Holzinger A. Usability engineering methods for software developers. Commun ACM 2005 Jan 01;48(1):71-74. [doi: 10.1145/1039539.1039541]
102. Nielsen J. Usability Engineering. Massachusetts, United States: Academic Press; 1993.
103. Abd-Alrazaq A, Bewick BM, Farragher T, Gardner P. Factors affecting patients' use of electronic personal health records in England: cross-sectional study. J Med Internet Res 2019 Jul 31;21(7):e12373 [FREE Full text] [doi: 10.2196/12373] [Medline: 31368442]
104. Razmak J, Bélanger C. Using the technology acceptance model to predict patient attitude toward personal health records in regional communities. Inf Technol People 2018 Apr 03;31(2):306-326. [doi: 10.1108/ITP-07-2016-0160]
105. Ukoha EP, Feinglass J, Yee LM. Disparities in electronic patient portal use in prenatal care: retrospective cohort study. J Med Internet Res 2019 Sep 23;21(9):e14445 [FREE Full text] [doi: 10.2196/14445] [Medline: 31586367]
106. Assadi V, Hassanein K. Consumer adoption of personal health record systems: a self-determination theory perspective. J Med Internet Res 2017 Jul 27;19(7):e270 [FREE Full text] [doi: 10.2196/jmir.7721] [Medline: 28751301]
107. Or CK, Karsh B, Severtson DJ, Burke LJ, Brown RL, Brennan PF. Factors affecting home care patients' acceptance of a web-based interactive self-management technology. J Am Med Inform Assoc 2011;18(1):51-59 [FREE Full text] [doi: 10.1136/jamia.2010.007336] [Medline: 21131605]
108. Or CK, Karsh B. A systematic review of patient acceptance of consumer health information technology. J Am Med Inform Assoc 2009;16(4):550-560 [FREE Full text] [doi: 10.1197/jamia.M2888] [Medline: 19390112]

109. Venkatesh V, Bala H. Technology acceptance model 3 and a research agenda on interventions. Decis Sci 2008 May;39(2):273-315. [doi: 10.1111/j.1540-5915.2008.00192.x]
110. Zulman DM, Nazi KM, Turvey CL, Wagner TH, Woods SS, An LC. Patient interest in sharing personal health record information: a web-based survey. Ann Intern Med 2011 Dec 20;155(12):805-810 [FREE Full text] [doi: 10.7326/0003-4819-155-12-201112200-00002] [Medline: 22184687]
111. DeVellis R. Scale Development: Theory and Applications. Thousand Oaks, CA: Sage Publication; 2003.
112. Polit DF, Beck CT, Owen SV. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Res Nurs Health 2007 Aug;30(4):459-467. [doi: 10.1002/nur.20199] [Medline: 17654487]
113. Streiner D, Norman G, Cairney J. Health Measurement Scales: A Practical Guide To Their Development And Use. Oxford, UK: Oxford University Press; 2014.
114. Dillman D. Mail and Internet Surveys: The Tailored Design Method. 2nd Edition. New York, NY: Wiley; 1999.
115. Andrews D, Nonnecke B, Preece J. Electronic survey methodology: a case study in reaching hard-to-involve internet users. Int J Hum Comput Interact 2003;16(2):185-210. [doi: 10.1207/s15327590ijhc1602_04]
116. Wright K. Researching internet-based populations: advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. J Comput Mediat Commun 2005 Apr 1;10(3):JCMC1034. [doi: 10.1111/j.1083-6101.2005.tb00259.x]
117. Chin W, Newsted P. Structural equation modeling analysis with small samples using partial least squares. In: Statistical Strategies For Small Sample Research. Thousand Oaks, CA: Sage Publications; 1999:307-341.
118. Chin W. Issues and opinion on structural equation modeling. MIS Q 1998;22(1):7-16. [doi: 10.5555/290231.290235]
119. Gefen D, Straub D, Boudreau M. Structural equation modeling and regression: guidelines for research practice. Commun Assoc Inf Syst 2000;4:1-77. [doi: 10.17705/1CAIS.00407]
120. Thomas D, Lu I, Cedzynski M. Partial least squares: a critical review and a potential alternative. Administrative Sciences Association of Canada (ASAC). 2005. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.1719&rep=rep1&type=pdf [accessed 2021-02-16]
121. Podsakoff PM, MacKenzie SB, Lee J, Podsakoff NP. Common method biases in behavioral research: a critical review of the literature and recommended remedies. J Appl Psychol 2003 Oct;88(5):879-903. [doi: 10.1037/0021-9010.88.5.879] [Medline: 14516251]
122. Pavlou P, Liang H, Xue Y. Understanding and mitigating uncertainty in online exchange relationships: a principal-agent perspective. MIS Q 2007 Mar;31(1):105-136. [doi: 10.2307/25148783]
123. Liang H, Saraf N, Hu Q, Xue Y. Assimilation of enterprise systems: the effect of institutional pressures and the mediating role of top management. MIS Q 2007 Mar;31(1):59-87. [doi: 10.2307/25148781]
124. Anderson JC, Gerbing DW. Structural equation modeling in practice: a review and recommended two-step approach. Psychol Bull 1988 May;103(3):411-423. [doi: 10.1037/0033-2909.103.3.411]
125. Ringle C, Wende S. SmartPLS 3. URL: http://www.smartpls.com [accessed 2021-05-12]
126. Lohmöller J. Latent Variable Path Modeling With Partial Least Squares. Heidelberg, Germany: Physica-Verlag Heidelberg; 1989.
127. Wilson B, Henseler J. Modeling reflective higher-order constructs using three approaches with PLS path modeling: a Monte Carlo comparison. University of Twente. 2007. URL: https://research.utwente.nl/en/publications/modeling-reflective-higher-order-constructs-using-three-approache [accessed 2021-02-18]
128. Fornell C, Larcker DF. Structural equation models with unobservable variables and measurement error: algebra and statistics. J Mark Res 1981 Aug;18(3):382-388. [doi: 10.2307/3150980]
129. Podsakoff PM, Organ DW. Self-reports in organizational research: problems and prospects. J Manag 1986 Dec 1;12(4):531-544. [doi: 10.1177/014920638601200408]
130. Tenenhaus M, Amato S, Vinzi V. A global goodness-of-fit index for PLS structural equation modelling. Proc XLII SIS Scient Meet 1 2004 Jan;1:739-742 [FREE Full text]
131. Tenenhaus M, Vinzi VE, Chatelin Y, Lauro C. PLS path modeling. Comput Stat Data Anal 2005 Jan 1;48(1):159-205. [doi: 10.1016/j.csda.2004.03.005]
132. Frank FR, Miller N. A Primer for Soft Modeling. Akron, Ohio: The University of Akron Press; 1992.
133. Wetzels M, Odekerken-Schröder G, van Oppen C. Using PLS path modeling for assessing hierarchical construct models: guidelines and empirical illustration. MIS Q 2009 Mar;33(1):177-195. [doi: 10.2307/20650284]
134. Warshaw PR. A new model for predicting behavioral intentions: an alternative to Fishbein. J Mark Res 1980 May;17(2):153-172. [doi: 10.1177/002224378001700201]
135. Lahteenmaki J, Leppanen J, Kaijanranta H. Interoperability of personal health records. Annu Int Conf IEEE Eng Med Biol Soc 2009;2009:1726-1729. [doi: 10.1109/IEMBS.2009.5333559] [Medline: 19964259]
136. Oliver RL. A cognitive model of the antecedents and consequences of satisfaction decisions. J Mark Res 1980 Nov;17(4):460-469 [FREE Full text] [doi: 10.2307/3150499]
137. Venkatesh V, Goyal S. Expectation disconfirmation and technology adoption: polynomial modeling and response surface analysis. MIS Q 2010 Jun;34(2):281-303. [doi: 10.2307/20721428]

138. Liu CF, Tsai YC, Jang FL. Patients' acceptance towards a web-based personal health record system: an empirical study in Taiwan. Int J Environ Res Public Health 2013 Oct 17;10(10):5191-5208 [FREE Full text] [doi: 10.3390/ijerph10105191] [Medline: 24142185]
139. Assadi V, Hassanein K. Continuance intention to use high maintenance information systems: the role of perceived maintenance effort. ECIS 2010 Proceedings. 2010. URL: https://aisel.aisnet.org/ecis2010/88/ [accessed 2021-03-15]
140. Cabitza F, Simone C, De Michelis G. User-driven prioritization of features for a prospective InterPersonal Health Record: perceptions from the Italian context. Comput Biol Med 2015 Apr;59:202-210. [doi: 10.1016/j.compbiomed.2014.03.009] [Medline: 24768267]
141. Ralston JD, Carrell D, Reid R, Anderson M, Moran M, Hereford J. Patient web services integrated with a shared medical record: patient use and satisfaction. J Am Med Inform Assoc 2007;14(6):798-806 [FREE Full text] [doi: 10.1197/jamia.M2302] [Medline: 17712090]
142. Gill G, Bhattacherjee A. Whom are we informing? Issues and recommendations for MIS research from an informing sciences perspective. MIS Q 2009 Jun;33(2):217-235. [doi: 10.2307/20650290]
143. Rosemann M, Vessey I. Toward improving the relevance of information systems research to practice: the role of applicability checks. MIS Q 2008;32(1):1-22. [doi: 10.2307/25148826]

Abbreviations

AVE: average variance extracted
CHI: consumer health informatics
EHR: electronic health record
EMR: electronic medical record
GoF: goodness of fit
HIS: health information system
IS: information system
PHR: personal health record
PLS: partial least squares
SEM: structural equation modeling
TAM: technology acceptance model

Edited by G Eysenbach; submitted 10.05.21; peer-reviewed by L Chirchir; comments to author 04.06.21; revised version received 07.06.21; accepted 01.08.21; published 24.09.21.

Please cite as:
Ruhi U, Majedi A, Chugh R
Sociotechnical Drivers and Barriers in the Consumer Adoption of Personal Health Records: Empirical Investigation
JMIR Med Inform 2021;9(9):e30322
URL: https://medinform.jmir.org/2021/9/e30322
doi:10.2196/30322
PMID:34343106

Original Paper

Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study

Lama Alsudias[1,2], BSc, MSc; Paul Rayson[1], BSc, PhD

[1]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

[2]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:
Lama Alsudias, BSc, MSc
School of Computing and Communications
Lancaster University
InfoLab21
Lancaster, LA1 4WA
United Kingdom
Phone: 44 1524 510357
Email: l.alsudias@lancaster.ac.uk

Abstract

Background: Twitter is a real-time messaging platform widely used by people and organizations to share information on many topics. Systematic monitoring of social media posts (infodemiology or infoveillance) could be useful to detect misinformation outbreaks as well as to reduce reporting lag time and to provide an independent complementary source of data compared with traditional surveillance approaches. However, such an analysis is currently not possible in the Arabic-speaking world owing to a lack of basic building blocks for research and dialectal variation.

Objective: We collected around 4000 Arabic tweets related to COVID-19 and influenza. We cleaned and labeled the tweets relative to the Arabic Infectious Diseases Ontology, which includes nonstandard terminology, as well as 11 core concepts and 21 relations. The aim of this study was to analyze Arabic tweets to estimate their usefulness for health surveillance, understand the impact of the informal terms in the analysis, show the effect of deep learning methods in the classification process, and identify the locations where the infection is spreading.

Methods: We applied the following multilabel classification techniques: binary relevance, classifier chains, label power set, adapted algorithm (multilabel adapted k-nearest neighbors [MLKNN]), support vector machine with naive Bayes features (NBSVM), bidirectional encoder representations from transformers (BERT), and AraBERT (transformer-based model for Arabic language understanding) to identify tweets appearing to be from infected individuals. We also used named entity recognition to predict the place names mentioned in the tweets.

Results: We achieved an F1 score of up to 88% in the influenza case study and 94% in the COVID-19 one. Adapting for nonstandard terminology and informal language helped to improve accuracy by as much as 15%, with an average improvement of 8%. Deep learning methods achieved an F1 score of up to 94% during the classifying process. Our geolocation detection algorithm had an average accuracy of 54% for predicting the location of users according to tweet content.

Conclusions: This study identified two Arabic social media data sets for monitoring tweets related to influenza and COVID-19. It demonstrated the importance of including informal terms, which are regularly used by social media users, in the analysis. It also proved that BERT achieves good results when used with new terms in COVID-19 tweets. Finally, the tweet content may contain useful information to determine the location of disease spread.

(JMIR Med Inform 2021;9(9):e27670) doi:10.2196/27670

KEYWORDS

Arabic; COVID-19; infectious disease; influenza; infodemiology; infoveillance; social listening; informal language; multilabel classification; natural language processing; named entity recognition; Twitter

Introduction

Background

Although millions of items of data appear every day on social media, artificial intelligence through natural language processing (NLP) and machine learning (ML) algorithms offers the chance to automate their analysis across many different areas, including health. In the area of health informatics and text mining, social media data, such as Twitter data, can be analyzed to calculate large-scale estimates of the number of infections and the spread of diseases, or help to predict epidemic events [1]; this field is known as infodemiology, and the systematic monitoring of social media posts and Internet information for public health purposes is known as infoveillance. However, previous research has focused almost exclusively on English data.

Time is clearly an important factor in the health surveillance domain. In other words, discovering infectious diseases as quickly as possible is beneficial for many organizations and populations, as we have seen internationally with COVID-19. It is also important to have multiple independent sources to corroborate evidence of the spread of infectious diseases.

Twitter is one of the main real-time platforms that can be used in health monitoring. However, it contains noisy and unrelated information; hence, there is a crucial need for information gathering, preprocessing, and filtering techniques to discard irrelevant information while retaining useful information. One key task is to differentiate between tweets written for different reasons where someone is infected or worried about a disease, taking into account the figurative usage of some words related to a disease or spread of infection [2].

While such tasks are obviously relevant globally, there is little previous research for Arabic-speaking countries. There are some characteristics of the Arabic language that make it more difficult to analyze compared with other languages, and NLP resources and methods are less well developed for Arabic than for English. Arabic, which has more than 26 dialects, is spoken by more than 400 million people around the world [3]. We hypothesize that Arabic speakers will use their own dialects in informal discourse when they express their pain, concerns, and feelings rather than using modern standard Arabic [4]. Table 1 describes some examples of Arabic words related to health that may represent different meanings owing to dialect differences. For instance, the word can be understood as influenza in Najdi dialect and feeling cold in Hejazi dialect [3].

Table 1. Some examples of Arabic words that have different meaning.

Word in Arabic	Potential meaning confusion sets
	Influenza (cold)/feeling cold
	Vaccination/reading supplication
	Runny nose/nosebleed
	Ointment/paint
	Sneezing (cold)/filter the liquid thing/be nominated for a position
	Antibiotic/opposite
	Tablets/pimples/some kind of food
	X-ray/sunlight
	Weakness/double
	Painkiller/home
	Prescription/method
	Medicine (like vitamin C fizz)/sparkling spring (fizz)

The real-world motivation of this work is to reduce the lag time and increase accuracy in detecting mentions of infectious diseases in order to support professional organizations in decreasing the spread, planning for medicine roll out, and increasing awareness in the general population. We also wish to show that Arabic tweets on Twitter can provide valuable data that may be used in the area of health monitoring by using informal, nonstandard, and dialectal language, which represents social media usage more accurately.

We focused on COVID-19 and influenza in particular owing to their rapid spread during seasonal epidemics or pandemics in the Arabic-speaking world and beyond. Most people recover within a week or two. However, young children, elderly people, and those with other serious underlying health conditions may experience severe complications, including infection, pneumonia, and death [5]. While it takes specialized medical knowledge to distinguish between the people infected by COVID-19 and influenza as the symptoms are similar, tracing and planning vaccination and isolation are important for both diseases. In addition, there may be some infected people who do not take the test because of personal concerns and lack of availability of tests in their city, or those who need support to self-isolate.

The overall question being answered in this paper is how NLP can improve the analysis of the spread of infectious diseases via social media. Our first main contribution is the creation of a new Arabic Twitter data set related to COVID-19 and influenza, which was labeled with 12 classes, including 11 originating from the Arabic Infectious Disease Ontology [6] and a new infection category. We used this ontology since there are no existing medical ontologies, such as International Classification of Diseases (ICD) and/or Systematized Nomenclature of Medicine-Clinical Terms (SNOMED), available that originate in Arabic [1]. Crucially, we also showed for the first time the usefulness of informal nonstandard disease-related terms using a multilabel classification methodology to find personal tweets related to COVID-19 or influenza in Arabic. We comparatively evaluated our results with and without the informal terms and showed the impact of including such terms in our study. Moreover, we showed the power of ML and deep learning algorithms in the classification process. Finally, we developed methods to identify the locations of the infectious disease spread using tweet content, and this also helped to inform dialect variants and choices.

Related Work

Previous studies have proven that NLP techniques can be used to analyze tweets for monitoring public health [7-12]. These studies have analyzed social media articles that support the surveillance of diseases in different languages such as Japanese, Chinese, and English. Diseases that were analyzed included listeria, influenza, swine flu, measles, meningitis, and others. As justified in the previous section, we will focus on previous work related to monitoring influenza and COVID-19 using Twitter data.

Influenza-Related Research

The Ailment Topic Aspect Model (ATAM) is a model designed by Paul and Dredze [13]. It uses Twitter messages to measure influenza rates in the United States. It was later extended to consider over a dozen ailments and apply several tasks such as syndromic surveillance and geographical disease monitoring. Similarly, an influenza corpus was created from Twitter [8]. The tweets needed to meet the following two conditions to include them in the training data with infected people and timing: (1) the person tweeting or a close contact is infected with the flu and (2) the tense should be the present tense or recent past tense.

The goal of a previous study [2] was to distinguish between flu tweets from infected individuals and others worried about infection in order to improve influenza surveillance. It applied multiple features in a supervised learning framework to find tweets indicating flu. Likewise, a sentiment analysis approach was used [14] to classify tweets that included 12 diseases, including influenza. A forecasting word model was designed [15] using several words, such as symptoms, that appear in tweets before epidemics to predict the number of patients infected with influenza.

A previous study [16] used unsupervised methods based on word embeddings to classify health-related tweets. The method achieved an accuracy of 87.1% for the classification of tweets being related or unrelated to a topic. Another study [17] concluded that there is a high correlation between flu tweets and Google Trends data.

A recent survey study [1] showed how ontologies may be useful in collecting data owing to the structured information they contain. However, there were serious challenges as medical ontologies may consist of medical terms, while the text itself may contain slang terms. The study suggested the inclusion of informal language from social media in the analysis process in order to improve the quality of epidemic intelligence in the future, but this was not implemented.

COVID-19–Related Research

Many researchers in computer science have made extensive efforts to show how they can help during pandemics. In terms of NLP and social media, there are various studies that support different languages with multiple goals. These goals include defining topics discussed in social media, detecting fake news, analyzing sentiments of tweets, and predicting the number of cases [18].

There have been multiple Arabic data sets published recently [19,20]. The authors explained the ways of collecting tweets, such as time period, keywords, and software library used in the search process, and summed up the statistics for the collected tweets. However, they only included statistical analysis and clustering to generate summaries with some suggestion of future work. Yet, there are some studies with specific goals, such as analysis of the reaction of citizens during a pandemic [21] and identification of the most frequent unigrams, bigrams, and trigrams of tweets related to COVID-19 [22]. In addition, considering the study by Alanazi et al [23] that identified the symptoms of COVID-19 from Arabic tweets, the authors noted the limitation that they used modern standard Arabic keywords only, and it would be important to consider dialectical keywords in order to better catch tweets on COVID-19 symptoms written in Arabic, because some Arab users post on social media in their own local dialect.

In a previous study, we analyzed COVID-19 tweets in the following three different ways: (1) identifying the topics discussed during the period, (2) detecting rumors, and (3) predicting the source of the tweets in order to investigate reliability and trust [24].

Critically, none of the above studies utilized the Arabic language for monitoring the spread of diseases. There are some Arabic studies that used Twitter with the goal of determining the correctness of health information [25], analyzing health services [26], and proving that Twitter is used by health professionals [27]. Moreover, other studies, which did not involve Arabic, used only formal language terminologies when collecting tweets, and we would argue that this is not representative of the language usage in social media posts.

Arabic Named Entity Recognition–Related Research

Previous research on named entity recognition (NER) aimed to accomplish the following two key goals: (1) the identification of named entities and (2) the classification of these entities, usually into coarse-grained categories, including personal names (PER), organizations (ORG), locations (LOC), and dates and times (DATE). In this study, our interest was in estimating one of these categories, which is the location element of the information on Twitter. NER methods use a variety of approaches, including rule-based, ML-based, deep learning–based, and hybrid approaches. These approaches can be used for Arabic, although specific issues arise, such as lack of capitalization, nominal confusability, agglutination, and absence of short vowels [28,29]. In addition, there are more challenges in terms of social media content, which includes Arabic dialects and informal terms. There is a lack of annotated data for NER in dialects. The application of NLP tools, originally designed for modern standard Arabic, on dialects leads to considerably less efficiency, and hence, we see the need to develop resources and tools specifically for Arabic dialects [29].

The goal of a previous study [30] was to illustrate a new approach for the geolocation of Arabic and English language tweets based on content by collecting contextual tweets. It proved that only 0.70% of users actually use the function of geospatial tagging of their own tweets; thus, other information should be used instead.

Data Collection and Filtering

There is a lack of an available and reliable Twitter corpora in Arabic in the health domain, which makes it necessary for us to create our own corpus. We obtained the data using the Twitter application programming interface (API) for the period between September 2019 and October 2020, and collected around 6 million tweets that contained influenza or COVID-19 keywords. The keywords are in the code that we will release on GitHub [31]. We collected the tweets weekly since the Twitter API does not otherwise allow us to retrieve enough historical tweets. We utilized keywords related to influenza and COVID-19 from the Arabic Infectious Diseases Ontology [6], which includes nonstandard terminology. We used a disease ontology because it has been shown to help in finding all the terms and synonyms related to the disease [14].

A previous survey [1] suggested the inclusion of informal text used in social media in medical ontologies and search processes when collecting data in order to improve the quality of epidemic intelligence. Therefore, we hypothesize that informal terms may help to find the relevant tweets related to diseases. Additionally, in the Arabic scenario, we hypothesize that we need to account for dialectal terms.

We filtered the tweets by excluding duplicates, advertisements, and spam. Using Python, we also cleaned the tweets by removing symbols, links, non-Arabic words, URLs, mentions, hashtags, numbers, and repeating characters. From the resulting data set, we took a sample of about 4000 unique tweets (2000 tweets on influenza and 2000 tweets on COVID-19). Then, we used a suite of approaches for preprocessing the tweets, applying the following processes in sequence: tokenization, normalization, and stop-word removal. Table 2 shows the number of tweets with each label from the ontology after filtering and preprocessing.

Table 2. The number of tweets in each label.

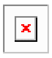
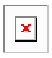
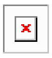
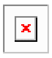
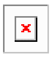
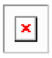
Label	Tweets[a], n	
	Influenza	COVID-19
Name of the disease	1544	1795
Slang term of the disease	456	327
Symptom	398	789
Cause	178	530
Prevention	666	209
Infection	51	15
Organ	2	202
Treatment	152	97
Diagnosis	25	2
Place of the disease spread	17	415
Infected category	52	12
Infected with	907	915

[a]Each tweet can have multiple labels.

Manual Coding

In order to create a gold standard corpus, our process started with tweet labeling by two Arabic native speakers, including the first author of the paper, following the guidelines of the annotation process described in Multimedia Appendix 1. We manually annotated each tweet with 1 or 0 to indicate Arabic Infectious Diseases Ontology classes, which are infectious disease name (ie, influenza and COVID-19 in our case), slang term, symptom, cause, prevention, infection, organ, treatment, diagnosis, place of disease spread, and infected category. We also labeled each tweet as 1 if the person who wrote the tweet was infected with influenza or COVID-19 and 0 if not. Table 3 describes some examples of Arabic influenza and COVID-19 tweets with their labels.

Table 3. Examples of tweets with their assigned labels (1 or 0).

Tweet in Arabic	Tweet in English	Name	Slang name	Symptom	Cause	Prevention	Infection	Organ	Treatment	Diagnosis	Place of disease spread	Infected category	Infected with
	What is the solution with flu, fever and cold killed me	1[a]	1	1	0	0	0	0	0	0	0	0	1[b]
	Influenza vaccination campaign in cooperation with King Khalid Hospital in Al-Kharj	1	0	0	0	1	0	0	0	0	0	0	0
	Flu morning	1	0	0	0	0	0	0	0	0	0	0	1
	When you have symptoms of a flu or cold, Does the clinic take a sample of nose and throat to check if its bacteria or a virus	0	1	1	1	0	0	1	0	1	0	0	0
	My experience with after my infection with the Covid-19 virus was confirmed, I did not initially care about eating food, enough water, and also food supplements, because the symptoms were slight, I noticed that the virus works in stages, at first I noticed sweating, headache, and then eye pain.	1	0	1	1	0	0	1	0	0	0	0	1
	Washing hands with soap and water, and wearing a medical mask ... Here are a number of precautionary measures that are still the best ways to prevent Corona	0	1	0	0	1	0	0	0	0	0	0	0

Tweet in Arabic	Tweet in English	Name	Slang name	Symptom	Cause	Prevention	Infection	Organ	Treatment	Diagnosis	Place of disease spread	Infected category	Infected with
	The first thing that struck me was lethargy, pain in the bones and muscles, a strange headache that was not painful but bothersome, and then had diarrhea. I did not expect Corona because the symptoms were mild, not like what people say. But I was sure when my sleep became strange, as if I woke up not asleep, and after that I fell asleep for an hour or two, and sometimes I did not sleep. .	1	0	1	0	0	0	1	0	0	0	0	1

[a]We labeled each tweet with 1 or 0 to indicate Arabic Infectious Diseases Ontology classes.

[b]We labeled each tweet as 1 if the person who wrote the tweet was infected and 0 if not.

Interrater Reliability

We used the Krippendorff alpha coefficient statistic, which supports multilabel input, to test the robustness of the classification scheme for both data sets [32]. The result showed that the Krippendorff alpha score was 0.84 in the influenza data set and 0.91 in the COVID-19 data set, which indicates strong agreement between the two manual coders. The remaining disagreement between the annotators was due to informal terms and Arabic dialects found in social media. For instance, can be understood as “cold is playing with us,” which represents that an uninfected person or flu is playing with us (indicating an infected person). Another example is , which in English means “get along with Corona is easier than the lockdown.” This may be classified as an infected person or an uninfected person because the word has various meanings.

Methods

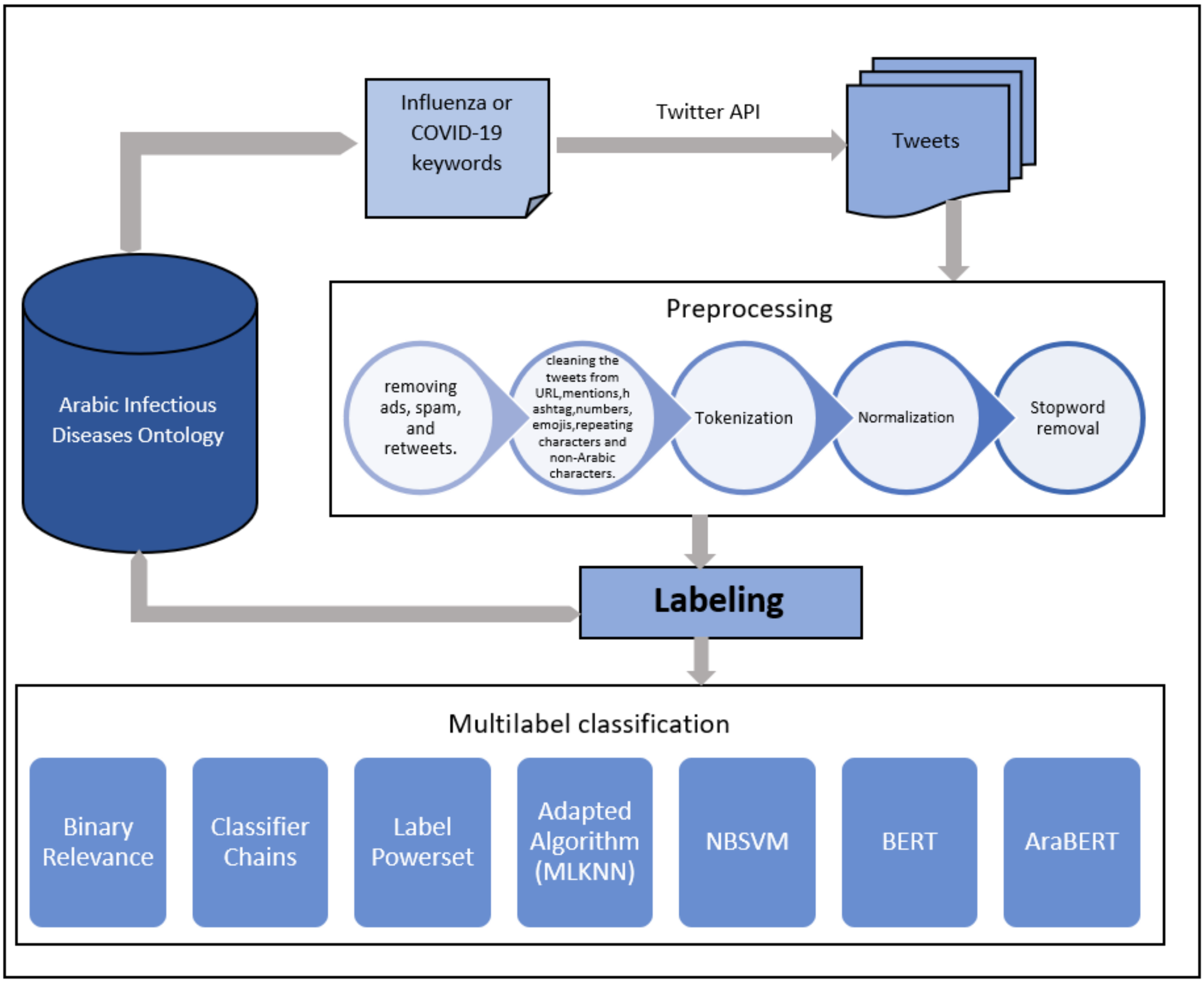
Overview

In order to create methods to find individuals who have been self-identified as infected and to determine their geolocation in the Twitter data set, we applied multiple supervised learning algorithms on the labeled data set and used NER on the tweet content.

Multilabel Classification

The overall architecture of our pipeline for finding infected people is shown in Figure 1. Using a supervised paradigm, we first annotated the corpus with labeling information as described above, before moving on to classify the tweets by applying machine and deep learning algorithms. We used this method for both the influenza and COVID-19 case studies. Each tweet has different labels assigned to it. For instance, the first example in Table 3 contains the labels influenza name (), slang term of influenza (), and symptom (). It also represents that the person is infected with influenza. Therefore, we assigned a value of 1 to these labels. On the other hand, the tweet does not include the labels cause, prevention, infection, organ, treatment, diagnosis, place of disease spread, and infected category. Thus, these were marked with 0.

Figure 1. System architecture. API: application programming interface; AraBERT: transformer-based model for Arabic language understanding; BERT: bidirectional encoder representations from transformers; MLKNN: multilabel adapted k-nearest neighbors; NBSVM: support vector machine with naive Bayes features.


Influenza or COVID-19 keywords
Twitter API
Tweets
Arabic Infectious Diseases Ontology
Preprocessing
removing ads, spam, and retweets.
cleaning the tweets from URL,mentions,h ashtag,numbers, emojis,repeating characters and non-Arabic characters.
Tokenization
Normalization
Stopword removal
Labeling
Multilabel classification
Binary Relevance
Classifier Chains
Label Powerset
Adapted Algorithm (MLKNN)
NBSVM
BERT
AraBERT

From Table 3, we can see that we have a multilabel classification problem where multiple labels are assigned to each tweet. Basically, the following three methods can be used to solve the problem: problem transformation, adapted algorithm, and ensemble approaches. For each method, there are different techniques that can be used. We applied the following algorithms, which represent ML and deep learning algorithms, to classify the tweets: (1) binary relevance, which treats each label as a separate single class classification problem; (2) classifier chains, which treats each label as a part of a conditioned chain of single-class classification problems, and it is useful to handle the class label relationships; (3) label power set, which transforms the problem into a multiclass problem with one multiclass classifier that is trained on all unique label combinations found in the training data; (4) adapted algorithm (MLKNN), which is a multilabel adapted k-nearest neighbors (KNN) classifier with Bayesian prior corrections; (5) support vector machine with naive Bayes features (NBSVM), which combines generative and discriminant models together by adding NB log-count ratio features to SVM [33]; (6) bidirectional encoder representations from transformers (BERT), which is a condition where all left and right meanings in both layers are used to pretrain deep bidirectional representations from unlabeled text [34]; and (7) transformer-based model for Arabic language understanding (AraBERT), which is a pretrained BERT model designed specifically for the Arabic language [35].

Since some labels were 0 for most tweets, we removed these labels in order to avoid overfitting. In other words, we removed the labels that did not appear in most tweets as shown in Table 3. The remaining important labels were determined depending on the disease case study because they represented different values for different tweets as justified in Table 2. For influenza, they are influenza name, slang term of influenza, symptom, prevention, treatment, and infected with. While for COVID-19, they are name, slang term of COVID-19, symptom, cause, place, and infected with. We also repeated the experiment twice to show the effectiveness of the informal terms in the results. One of them had the labels “disease name,” “slang term of infectious disease,” and “infected with,” and the other had all labels, except “slang term of infectious disease” in both case studies.

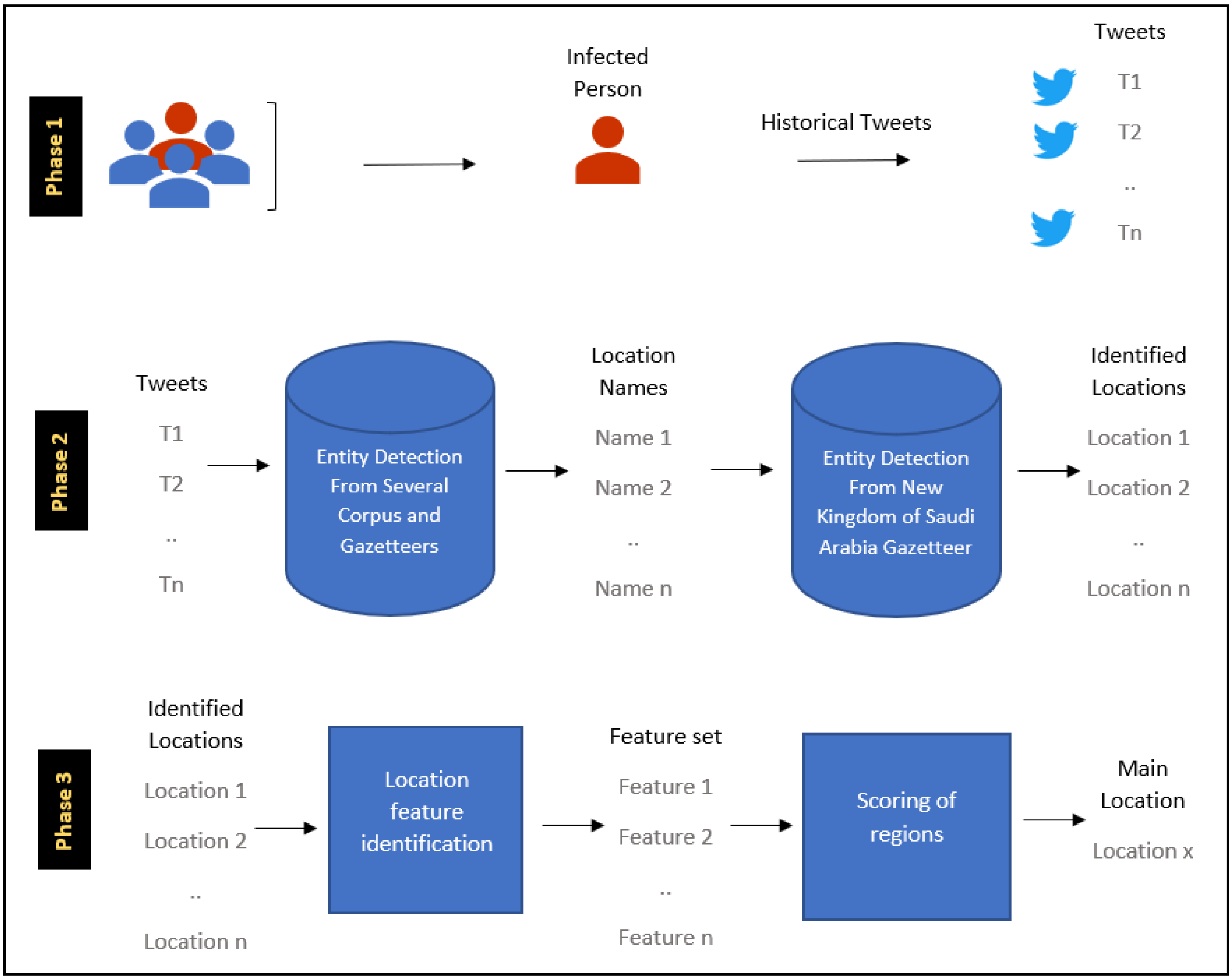
In our study, we used the Python scikit-multilearn [36] and ktrain [37] libraries and applied different models. To extract the features from the processed training data, we used a word frequency approach. We split the entire sample into 75% training and 25% testing sets.

NER

We followed NER systems that used ML algorithms to learn NE tag decisions from annotated text. We used the conditional random fields (CRF) algorithm because it achieved better results than other supervised NER ML techniques in previous studies [29].

There were three phases in our geolocation detection algorithm as shown in Figure 2. In phase 1, the infected person was specified from the multilabel classification algorithm described in the previous section. Then, we retrieved the historical tweets of this person (around 3000 tweets per person on average) and passed them to the next phase.

Figure 2. Three phases of the geolocation detection algorithm.



Phase 2 consisted of two consecutive stages. First, the tweets were submitted to a named entity detection algorithm to select location records from multiple corpora and gazetteers, including ANERCorp [38,39], and ANERGazet [40]. A set of location names needs to be filtered out from the general names and ambiguous ones. For example, the word (Bali in English) can be a province in Indonesia or "my mind" as an informal term in Arabic. This step is important in order to ensure that all unrelated location names are not included in the final phase. Second, the identified locations were determined by applying our new entity detection gazetteer, which represents Saudi Arabia regions, cities, and district. The data, which will be released on GitHub [31], are public data collected from the Saudi Post website [41].

In phase 3, common features were identified, such as the most frequent locations, as well as other features, such as occurrence time, which gives a higher score for locations within the last 6 months. Then, each location is scored by a number, which allows us to rank the list and determine the best estimated main location of the user.

After each tweet set with a predictable location, we compared this location with the location field mentioned in the user account, which is not always set by the user because it is an optional field. Here, we kept only users with valuable location information in either the location or description fields.

Ethical Considerations

Although Twitter has obtained informed consent from users to share information, there was a need to obtain research ethics approval from our university, especially considering our focus on health-related topics [42]. Ethical approval for this study was obtained from Lancaster University on June 21, 2019 [43].

Results

Multilabel Classification

A multilabel classification problem is more complex than binary and multiclass classification problems. Therefore, various performance measures were calculated to evaluate the classification process, such as accuracy, F1 score, recall, precision, area under the receiver operating characteristic curve (AUC), and Hamming loss [44]. For all these measures, except Hamming loss, higher scores are better. For Hamming loss, smaller values reflect better performance. It is important to note that the accuracy score function in multilabel classification computes only subset accuracy, which means a sample of labels will be taken in the calculation process, as mentioned previously [36].

Table 4 illustrates the performance measures of the seven models on our training data set with six, five, and three labels for the influenza case study. In the six labels, which are “influenza name,” “slang term of influenza,” “symptom,” “prevention,” “treatment,” and “infected with,” the classifier chains algorithm achieved the highest results in most measures compared with the other algorithms. It had an F1 score of 86.1%, recall of 81.0%, precision of 91.8%, AUC of 88.6%, accuracy of 56.2%, and Hamming loss of 8.9%. The label power set algorithm provided a result slightly lower than the classifier chain by around 2%. The lowest F1 score was observed for NBSVM, which was 58.9%.

The repeated experiment results for the seven models on our training data set with three labels, which were “influenza name,” “slang term of influenza,” and “infected with,” and five labels, which were “influenza name,” “symptom,” “prevention,” “treatment,” and “infected with,” are described in Table 4. There was up to 20% enhancement for accuracy in the seven algorithms. The highest F1 score was achieved by the classifier chains algorithm, which was 88.8%. The recall and precision ranged from 60% to 92%. Consequently, informal terms were shown to represent key factors in the classification process.

Table 5 shows the performance measures of the seven models on our training data set with six, five, and three labels for the COVID-19 case study. Here, the six labels were different from those in the previous case study because they were determined according to the results from the number of tweets in each label as explained in Table 2. The six labels were “COVID-19 name,” “slang term of COVID-19,” “symptom,” “cause,” “place of disease spread,” and “infected with category.” The best results were achieved by the BERT algorithm with an F1 score of 88.2%, recall of 86.7%, precision of 89.7%, AUC of 90.3%, accuracy of 62.0%, and Hamming loss of 8.8%.

The repeated experiment results for the seven models on our training data set with three labels, which were “COVID-19 name,” “slang term of COVID-19,” and “infected with,” and five labels, which were “COVID-19 name,” “symptom,” “cause,” “place of disease spread,” and “infected with category” are described in Table 5. There was up to 20% enhancement for accuracy in the seven algorithms. The highest F1 score was achieved by the BERT algorithm, which was 94.8%, followed by AraBERT, which was 93.3%. The informal terms in the COVID-19 case study showed around 15% enhancement in the evaluation results.

Table 4. Training results of the seven algorithms with six, five, and three labels for the influenza case study.

Number of labels and multilabel classification techniques	F1 score (%)	Recall (%)	Precision (%)	AUC[a] (%)	Accuracy (%)	Hamming loss (%)
Six[b]						
Binary relevance	73.1	74.4	71.9	79.7	39.6	18.7
Classifier chains	86.1	81.0	91.8	88.6	56.2	8.9
Label power set	85.7	83.8	87.6	88.7	56.2	9.7
Adapted algorithm (MLKNN[c])	76.9	75.5	78.4	82.3	39.9	15.5
BERT[d]	78.1	83.4	73.4	85.4	38.9	13.7
AraBERT[e]	79.7	72.7	88.2	83.9	49.2	12.5
NBSVM[f]	58.9	46.3	81.2	70.9	26.8	18.9
Five[g]						
Binary relevance	75.5	76.9	74.1	80.7	45.1	18.3
Classifier chains	88.0	85.7	90.5	90.2	64.9	8.5
Label power set	87.6	86.2	89.2	90.0	63.9	8.9
Adapted algorithm (MLKNN)	79.9	76.4	83.9	84.0	47.9	14.0
BERT	84.1	83.1	85.0	88.0	57.5	10.3
AraBERT	87.3	86.3	88.4	90.0	64.3	9.0
NBSVM	61.6	49.7	81.2	72.0	26.8	20.2
Three[h]						
Binary relevance	80.8	80.0	81.7	81.2	60.4	18.8
Classifier chains	88.8	85.7	92.2	89.3	72.4	10.7
Label power set	88.3	88.0	88.6	88.4	70.8	11.6
Adapted algorithm (MLKNN)	80.9	84.7	77.5	80.2	54.0	19.8
BERT	87.6	93.9	82.1	88.9	68.1	11.7
AraBERT	85.9	81.5	90.9	86.8	66.9	13.1
NBSVM	79.5	75.1	84.3	82.1	59.9	17.1

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

[b] The six labels are “influenza name,” “slang term of influenza,” “symptom,” “prevention,” “treatment,” and “infected with.”

[c] MLKNN: multilabel adapted k-nearest neighbors.

[d] BERT: bidirectional encoder representations from transformers.

[e] AraBERT: transformer-based model for Arabic language understanding.

[f] NBSVM: support vector machine with naive Bayes features.

[g] The five labels are “influenza name,” “symptom,” “prevention,” “treatment,” and “infected with.”

[h] The three labels are “influenza name,” “slang term of influenza,” and “infected with.”

Table 5. Training results of the seven algorithms with six, five, and three labels for the COVID-19 case study.

Number of labels and multilabel classification technique	F1 score (%)	Recall (%)	Precision (%)	AUC[a] (%)	Accuracy (%)	Hamming loss (%)
Six[b]						
Binary relevance	54.6	52.8	56.6	64.0	15.6	33.3
Classifier chains	53.9	49.8	58.7	64.2	18.5	32.3
Label power set	58.6	59.4	57.9	66.5	22.2	31.8
Adapted algorithm (MLKNN[c])	54.5	51.0	58.4	64.4	10.0	32.4
BERT[d]	88.2	86.7	89.7	90.3	62.0	8.8
AraBERT[e]	82.0	84.4	79.8	86.0	50.5	13.6
NBSVM[f]	64.3	51.7	85.0	73.1	20.7	21.7
Five[g]						
Binary relevance	57.0	56.0	58.1	63.1	15.8	35.9
Classifier chains	56.2	53.0	59.9	63.3	18.3	35.1
Label power set	60.8	63.4	58.4	65.0	22.0	34.8
Adapted algorithm (MLKNN)	56.5	54.6	58.7	63.1	10.4	35.7
BERT	87.3	87.9	86.7	88.9	59.0	10.9
AraBERT	86.3	92.7	80.7	88.6	53.9	12.1
NBSVM	55.2	40.6	86.4	67.9	17.9	28.0
Three[h]						
Binary relevance	68.5	69.0	68.0	69.2	36.9	30.8
Classifier chains	69.7	68.1	71.4	71.2	39.9	28.7
Label power set	70.3	69.0	71.5	71.6	40.1	28.3
Adapted algorithm (MLKNN)	71.6	70.7	72.6	72.8	41.4	27.1
BERT	94.8	96.4	93.3	94.9	93.2	5.1
AraBERT	93.3	94.8	91.9	93.5	85.3	6.5
NBSVM	70.6	59.6	86.5	75.4	46.5	24.2

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

[b] The six labels are "COVID-19 name," "slang term of COVID-19," "symptom," "cause," "place of the disease spread," and "infected with category."

[c] MLKNN: multilabel adapted k-nearest neighbors.

[d] BERT: bidirectional encoder representations from transformers.

[e] AraBERT: transformer-based model for Arabic language understanding.

[f] NBSVM: support vector machine with naive Bayes features.

[g] The five labels are "COVID-19 name," "symptom," "cause," "place of the disease spread," and "infected with category."

[h] The three labels are "COVID-19 name," "slang term of COVID-19," and "infected with."

NER

A key point to be noted is that our geolocation detection evaluation is based on the location of users where they were tweeting. We filtered tweets that did not have any information in the location field and/or had nonplausible locations, such as moon and space. We created a manually annotated set from the information in the location field in order to demonstrate greater accuracy. This is due to the ambiguous information in the location field that can be detected by hand. For instance, we found some adjectives of the location, like ☒☒☒ and ☒, referring to Jeddah city in Saudi Arabia.

In the influenza study, around 907 users were classified as infected with influenza, and 397 of these users provided valuable information in their accounts that could be used to identify the location. As a result, our algorithm achieved an accuracy of 45.8% for predicting locations.

Regarding the COVID-19 study, 915 people were considered to be infected, and around 358 user accounts had useful information about the location. Therefore, after applying the algorithm, the accuracy was up to 63.6% for identifying the locations of the infected users.

Discussion

Principal Findings

To understand the effect of deep learning algorithms on the classification process, we needed to compare the results of the ML algorithms with deep learning ones in the two case studies for influenza and COVID-19. In the influenza study, the results of deep learning algorithms and ML ones were close to each other. In other words, there was no improvement in the results when applying deep learning methods, such as BERT and AraBERT. On the other hand, in the COVID-19 case study, there was up to a 25% enhancement in the results when applying BERT and/or AraBERT. These results helped to confirm that deep learning methods show good returns when dealing with new terms or unknown vocabularies that represent COVID-19 terms.

By applying our previous work [45], which classified the sources of the tweets into the following five types: academic, media, government, health professional, and public, we found that informal language was used in the public type (examples 1, 3, and 7 in Table 3), while the other types (academic, media, government, and health professional) utilized more formal styles (examples 2, 4, 5, and 6 in Table 3). Hence, disease-related slang names or other symptoms play an important role in detecting the disease mentions in social media. People not only used slang terms but also expressed their feelings using other terms such as metaphors [46]. For example, "☒," which means "hi flu," shows that the person, who wrote the tweet, was affected by flu. Here, 71.9% of the tweets proved that there was a relationship among the informal language used by flu-infected people.

We also found that there was a relationship among the "symptom," "prevention," and "infected with" labels. Overall, 64.3% of people infected by influenza sent tweets mentioning symptoms, such as sneezing, headache, coughing, and fever. Among tweets about prevention, 69.3% were written by a person who was not infected with influenza. However, there were a number of tweets that broke these patterns. In other words, we observed tweets written about symptoms that did not represent an infected person or tweets written about prevention that represented an infected person. Table 6 shows some examples of the tweets that described these relationships.

Table 6. Examples of tweets describing the relationships among the symptom, prevention, place, and infected with labels.

Tweet in Arabic	Tweet in English	Description
☒	Flu headache is bad	The relationship between symptom and infected with influenza
☒	I think I will die from flu; I sneeze 10 times from the time I wake up	The relationship between symptom and infected with influenza
☒	The flu vaccine does not prevent colds, as some believe, but it prevents serious influenza A and B infections that kill large numbers around the world	The relationship between prevention and noninfected with influenza
☒	Corona, what did you do for me? For two weeks, I will not be able to feel the taste of something	The relationship between symptom and infected with COVID-19
☒	Riyadh records 320 new coronavirus cases and 15 deaths	The relationship between place and noninfected with COVID-19
☒	Adhere to the precautions and prevention from Corona, as the wave has really started, so wear masks, stay away from gatherings, and sterilize and wash your hands with soap and water for a period of no less than thirty seconds	The relationship between prevention and noninfected with COVID-19

The study by Saker et al [47], which was published recently, proved that users who tested positive for COVID-19 also reported their symptoms using Twitter. Alanazi et al [23] described the most common COVID-19 symptoms from Arabic tweets in their study. These symptoms can be further evaluated in clinical settings and used in a COVID-19 risk estimate in near real time.

There are many ways to know the location of the Twitter user, such as geocoordinates, place field, user location, and tweet content. The most accurate method is using the network geolocation system for either the tweet or the user. However, because it is an optional field, less than 3% of users provide this information [19,48]. In addition, there is noisy information in the user location field because users can type anything like "home" or "in the heart of my dad." As a result, we used the tweet content by assuming that users mentioned helpful information when they tweeted.

On the other hand, some researchers have tried to predict the location of the user using dialect identification from the tweet content [49]. Although this may prove fruitful, in our scenario, it may not reflect the current location that would be required, since a person may tweet in the Egyptian dialect but live in Saudi Arabia.

Conclusion

This paper has, for the first time, shown that Arabic social media data contain a variety of suitable information for monitoring influenza and COVID-19, and crucially, it has improved on previous research methodologies by including informal language and nonstandard terminology from social media, which have been shown to help in filtering unrelated tweets. It should be noted that we are not trying to provide a single source of information for public health bodies to use, but want to provide a comparable information source through which to triangulate and corroborate estimates of disease spread against other more traditional sources.

We also introduced a new Arabic social media data set for analyzing tweets related to influenza and COVID-19. We labeled the tweets for categories in the Arabic Infectious Disease Ontology, which includes nonstandard terminology. Then, we used multilabel classification techniques to replicate the manual classification. The results showed a high F1 score for the classification task and showed how nonstandard terminology and informal language are important in the classification process, with an average improvement of 8.8%. The data set, including tweet IDs, manually assigned labels, and other resources used in this paper, have been released freely for academic research purposes, with a DOI via Lancaster University's research portal [50].

Moreover, we applied an NER algorithm on the tweet content to determine the location and spread of infection. Although the number of users was limited, the results showed good accuracy in the analysis process.

There are several further directions to enhance the performance of the system in the future, including expanding the data used to train the classifier, analyzing different infectious diseases, and using more NLP techniques and linguistic features.

Acknowledgments

The authors thank Nouran Khallaf, who is a PhD student at Leeds University (mlnak@leeds.ac.uk), for her help in labeling the tweets. LA wishes to thank King Saud University for funding her PhD study.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Guidelines for annotating tweets.
[DOCX File , 22 KB - medinform_v9i9e27670_app1.docx]

References

1. Joshi A, Karimi S, Sparks R, Paris C, Macintyre CR. Survey of Text-based Epidemic Intelligence. ACM Comput. Surv 2020 Jan 21;52(6):1-19. [doi: 10.1145/3361141]
2. Lamb A, Paul M, Dredze M. Separating Fact from Fear: Tracking Flu Infections on Twitter. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013 Presented at: 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2013; Atlanta, GA, USA p. 789-795.
3. Versteegh K. The Arabic Language. Edinburgh, UK: Edinburgh University Press; 2014.
4. Hadziabdic E, Hjelm K. Arabic-speaking migrants' experiences of the use of interpreters in healthcare: a qualitative explorative study. Int J Equity Health 2014 Jun 16;13(1):49-12 [FREE Full text] [doi: 10.1186/1475-9276-13-49] [Medline: 24934755]
5. World Health Organization. URL: https://www.who.int/ [accessed 2020-03-01]
6. Alsudias L, Rayson P. Developing an Arabic Infectious Disease Ontology to Include Non-Standard Terminology. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020 Presented at: 12th Language Resources and Evaluation Conference; May 2020; Marseille, France p. 4842-4850 URL: https://aclanthology.org/2020.lrec-1.596/
7. Paul M, Dredze M. You Are What You Tweet: Analyzing Twitter for Public Health. Proceedings of the International AAAI Conference on Web and Social Media 2021;5(1):265-272. [doi: 10.1145/2405716.2405728]
8. Aramaki E, Maskawa S, Morita M. Twitter catches the flu: detecting influenza epidemics using Twitter. In: EMNLP '11: Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2011 Presented at: Conference on Empirical Methods in Natural Language Processing; July 27-31, 2011; Edinburgh, UK p. 1568-1576 URL: https://dl.acm.org/doi/10.5555/2145432.2145600 [doi: 10.5555/2145432.2145600]
9. Breland JY, Quintiliani LM, Schneider KL, May CN, Pagoto S. Social Media as a Tool to Increase the Impact of Public Health Research. Am J Public Health 2017 Dec;107(12):1890-1891. [doi: 10.2105/AJPH.2017.304098] [Medline: 29116846]
10. Sinnenberg L, Buttenheim AM, Padrez K, Mancheno C, Ungar L, Merchant RM. Twitter as a Tool for Health Research: A Systematic Review. Am J Public Health 2017 Jan;107(1):e1-e8. [doi: 10.2105/AJPH.2016.303512] [Medline: 27854532]
11. Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EHY, Olsen JM, et al. Using Social Media for Actionable Disease Surveillance and Outbreak Management: A Systematic Literature Review. PLoS One 2015 Oct 5;10(10):e0139701 [FREE Full text] [doi: 10.1371/journal.pone.0139701] [Medline: 26437454]

12. Paul M, Sarker A, Brownstein J, Nikfarjam A, Scotch M, Smith K, et al. Social media mining for public health monitoring and surveillance. 2016 Presented at: Pacific Symposium on Biocomputing; 2016; Hawaii, USA p. 468-479. [doi: 10.1142/9789814749411_0043]
13. Paul M, Dredze M. A model for mining public health topics from Twitter. Health 2012;11:16.
14. Ji X, Chun S, Geller J. Knowledge-Based Tweet Classification for Disease Sentiment Monitoring. In: Pedrycz W, Chen S, editors. Sentiment Analysis and Ontology Engineering. Studies in Computational Intelligence, vol 639. Cham: Springer; 2016:425-454.
15. Iso H, Wakamiya S, Aramaki E. Forecasting Word Model: Twitter-based Influenza Surveillance and Prediction. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016 Presented at: 26th International Conference on Computational Linguistics: Technical Papers; December 2016; Osaka, Japan p. 76-86 URL: https://aclanthology.org/C16-1008/
16. Dai X, Bikdash M, Meyer B. From social media to public health surveillance: Word embedding based clustering method for twitter classification. 2017 Presented at: SoutheastCon 2017; March 30-April 2, 2017; Concord, NC, USA p. 1-7. [doi: 10.1109/secon.2017.7925400]
17. Hong Y, Sinnott R. A Social Media Platform for Infectious Disease Analytics. In: Gervasi O, editor. Computational Science and Its Applications – ICCSA 2018. ICCSA 2018. Lecture Notes in Computer Science, vol 10960. Cham: Springer; 2018:526-540.
18. Chandrasekaran R, Mehta V, Valkunde T, Moustakas E. Topics, Trends, and Sentiments of Tweets About the COVID-19 Pandemic: Temporal Infoveillance Study. J Med Internet Res 2020 Oct 23;22(10):e22624 [FREE Full text] [doi: 10.2196/22624] [Medline: 33006937]
19. Qazi O, Imran M, Ofli F. GeoCoV19. SIGSPATIAL Special 2020 Jun 05;12(1):6-15. [doi: 10.1145/3404820.3404823]
20. Shuja J, Alanazi E, Alasmary W, Alashaikh A. Covid-19 open source data sets: A comprehensive survey. Applied Intelligence 2020:1-30. [doi: 10.1101/2020.05.19.20107532]
21. Addawood A, Alsuwailem A, Alohali A, Alajaji D, Alturki M, Alsuhaibani J, et al. Tracking and understanding public reaction during COVID-19: Saudi Arabia as a use case. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. 2020 Presented at: 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020; November 20, 2020; Online. [doi: 10.18653/v1/2020.nlpcovid19-2.24]
22. Hamoui B, Alashaikh A, Alanazi E. What Are COVID-19 Arabic Tweeters Talking About? In: Chellappan S, Choo K, Phan N, editors. Computational Data and Social Networks. Cham: Springer International Publishing; 2020:425-436.
23. Alanazi E, Alashaikh A, Alqurashi S, Alanazi A. Identifying and Ranking Common COVID-19 Symptoms From Tweets in Arabic: Content Analysis. J Med Internet Res 2020 Nov 18;22(11):e21329 [FREE Full text] [doi: 10.2196/21329] [Medline: 33119539]
24. Alsudias L, Rayson P. COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. 2020 Presented at: 1st Workshop on NLP for COVID-19 at ACL 2020; July 2020; Online URL: https://aclanthology.org/2020.nlpcovid19-acl.16/
25. Alnemer K, Alhuzaim W, Alnemer A, Alharbi B, Bawazir A, Barayyan O, et al. Are Health-Related Tweets Evidence Based? Review and Analysis of Health-Related Tweets on Twitter. J Med Internet Res 2015 Oct 29;17(10):e246 [FREE Full text] [doi: 10.2196/jmir.4898] [Medline: 26515535]
26. Alayba A, Palade V, England M, Iqbal R. Arabic language sentiment analysis on health services. 2017 Presented at: 1st International Workshop on Arabic Script Analysis and Recognition (ASAR); April 3-5, 2017; Nancy, France p. 114-118. [doi: 10.1109/asar.2017.8067771]
27. Alsobayel H. Use of Social Media for Professional Development by Health Care Professionals: A Cross-Sectional Web-Based Survey. JMIR Med Educ 2016 Sep 12;2(2):e15 [FREE Full text] [doi: 10.2196/mededu.6232] [Medline: 27731855]
28. Shaalan K, Oudah M. A hybrid approach to Arabic named entity recognition. Journal of Information Science 2013 Oct 16;40(1):67-87. [doi: 10.1177/0165551513502417]
29. Zirikly A, Diab M. Named Entity Recognition for Arabic Social Media. In: Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing. 2015 Presented at: 1st Workshop on Vector Space Modeling for Natural Language Processing; June 2015; Denver, CO, USA p. 176-185. [doi: 10.3115/v1/w15-1524]
30. Khanwalkar S, Seldin M, Srivastava A, Kumar A, Colbath S. Content-based geo-location detection for placing tweets pertaining to trending news on map. 2013 Presented at: Fourth International Workshop on Mining Ubiquitous and Social Environments; 2013; Prague, Czech Republic.
31. Lama Alsudias. GitHub. URL: https://github.com/alsudias [accessed 2021-08-27]
32. Artstein R, Poesio M. Inter-Coder Agreement for Computational Linguistics. Computational Linguistics 2008 Dec;34(4):555-596. [doi: 10.1162/coli.07-034-r2]
33. Wang S, Manning C. Baselines and Bigrams: Simple, Good Sentiment and Topic Classification. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2012 Presented at: 50th Annual Meeting of the Association for Computational Linguistics; July 2012; Jeju Island, Korea p. 90-94 URL: https://aclanthology.org/P12-2018/

34. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. Preprint posted online May 24, 2019 [FREE Full text]
35. Antoun W, Baly F, Hajj H. AraBERT: Transformer-based Model for Arabic Language Understanding. In: Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection. 2020 Presented at: 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection; May 2020; Marseille, France p. 9-15.
36. Szymanski P, Kajdanowicz T. A scikit-based Python environment for performing multi-label classification. arXiv. Preprint posted online December 10, 2018 [FREE Full text]
37. Maiya A. ktrain: A Low-Code Library for Augmented Machine Learning. arXiv. Preprint posted online July 31, 2020 [FREE Full text]
38. Benajiba Y, Rosso P. Arabic named entity recognition using conditional random fields. 2008 Presented at: Workshop on HLT & NLP within the Arabic World, LREC; 2008; Citeseer p. 143-153 URL: http://personales.upv.es/prosso/resources/BenajibaRosso_LREC08.pdf
39. Obeid O, Zalmout N, Khalifa S, Taji D, Oudah M, Alhafni B, et al. CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing. In: Proceedings of the 12th Language Resources and Evaluation Conference. 2020 Presented at: 12th Language Resources and Evaluation Conference; May 2020; Marseille, France p. 7022-7032.
40. Benajiba Y, Rosso P, BenedíRuiz J. ANERsys: An Arabic Named Entity Recognition System Based on Maximum Entropy. In: Gelbukh A, editor. Computational Linguistics and Intelligent Text Processing. CICLing 2007. Lecture Notes in Computer Science, vol 4394. Berlin, Heidelberg: Springer; 2007:143-153.
41. National Address Maps. URL: https://maps.splonline.com.sa/ [accessed 2021-08-27]
42. Ahmed W, Bath P, Demartini G. Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In: Woodfield K, editor. The Ethics of Online Research (Advances in Research Ethics and Integrity, Vol. 2). Bingley, UK: Emerald Publishing Limited; 2017:79-107.
43. Research Ethics. Lancaster University. URL: https://www.lancaster.ac.uk/sci-tech/research/ethics [accessed 2019-06-01]
44. Wu X, Zhou Z. A unified view of multi-label performance measures. In: ICML'17: Proceedings of the 34th International Conference on Machine Learning. 2017 Presented at: 34th International Conference on Machine Learning; August 6-11, 2017; Sydney, NSW, Australia p. 3780-3788 URL: https://dl.acm.org/doi/10.5555/3305890.3306072
45. Alsudias L, Rayson P. Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases. 2019 Presented at: 3rd Workshop on Arabic Corpus Linguistics; July 22, 2019; Cardiff, United Kingdom p. 22-30 URL: https://aclanthology.org/W19-5604.pdf
46. Semino E, Demjén Z, Demmen J, Koller V, Payne S, Hardie A, et al. The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. BMJ Support Palliat Care 2017 Mar 05;7(1):60-66 [FREE Full text] [doi: 10.1136/bmjspcare-2014-000785] [Medline: 25743439]
47. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi M, Yang Y. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc 2020 Aug 01;27(8):1310-1315 [FREE Full text] [doi: 10.1093/jamia/ocaa116] [Medline: 32620975]
48. Dredze M, Paul M, Bergsma S, Tran H. Carmen: A twitter geolocation system with applications to public health. 2013 Presented at: AAAI workshop on expanding the boundaries of health informatics using AI (HIAI); 2013; Citeseer.
49. Abdul-Mageed M, Zhang C, Bouamor H, Habash N. NADI 2020: The First Nuanced Arabic Dialect Identification Shared Task. In: Proceedings of the Fifth Arabic Natural Language Processing Workshop. 2020 Presented at: Fifth Arabic Natural Language Processing Workshop; December 12, 2020; Barcelona, Spain (Online) p. 97-110 URL: https://aclanthology.org/2020.wanlp-1.9.pdf
50. Lama Alsudias. Research Portal | Lancaster University. 2021. URL: https://www.research.lancs.ac.uk/portal/en/people/lama-alsudias(2b6a561a-ef0f-4058-a713-c454fb133694)/datasets.html [accessed 2021-02-01]

Abbreviations

API: application programming interface
AraBERT: transformer-based model for Arabic language understanding
AUC: area under the receiver operating characteristic curve
BERT: bidirectional encoder representations from transformers
ML: machine learning
MLKNN: multilabel adapted k-nearest neighbors
NBSVM: support vector machine with naive Bayes features
NER: named entity recognition
NLP: natural language processing

Edited by C Lovis; submitted 02.02.21; peer-reviewed by S Doan, D Huang; comments to author 06.04.21; revised version received 20.04.21; accepted 20.06.21; published 17.09.21.

Please cite as:
Alsudias L, Rayson P
Social Media Monitoring of the COVID-19 Pandemic and Influenza Epidemic With Adaptation for Informal Language in Arabic Twitter Data: Qualitative Study
JMIR Med Inform 2021;9(9):e27670
URL: https://medinform.jmir.org/2021/9/e27670
doi:10.2196/27670
PMID:34346892

Original Paper

Automatic Classification of Thyroid Findings Using Static and Contextualized Ensemble Natural Language Processing Systems: Development Study

Dongyup Shin[1], MSc; Hye Jin Kam[2], PhD; Min-Seok Jeon[3], BSc; Ha Young Kim[1], PhD

[1]Graduate School of Information, Yonsei University, Seoul, Republic of Korea

[2]Healthcare, Life Solution Cluster, New Business Unit, Hanwha Life Insurance Co Ltd, Seoul, Republic of Korea

[3]Data Analysis Team, Aimmed Co Ltd, Seoul, Republic of Korea

Corresponding Author:
Ha Young Kim, PhD
Graduate School of Information
Yonsei University
New millennium hall 420, Yonsei-ro 50
Seodaemun-gu
Seoul, 03722
Republic of Korea
Phone: 82 10 4094 2392
Email: hayoung.kim@yonsei.ac.kr

Abstract

Background: In the case of Korean institutions and enterprises that collect nonstandardized and nonunified formats of electronic medical examination results from multiple medical institutions, a group of experienced nurses who can understand the results and related contexts initially classified the reports manually. The classification guidelines were established by years of workers' clinical experiences and there were attempts to automate the classification work. However, there have been problems in which rule-based algorithms or human labor–intensive efforts can be time-consuming or limited owing to high potential errors. We investigated natural language processing (NLP) architectures and proposed ensemble models to create automated classifiers.

Objective: This study aimed to develop practical deep learning models with electronic medical records from 284 health care institutions and open-source corpus data sets for automatically classifying 3 thyroid conditions: healthy, caution required, and critical. The primary goal is to increase the overall accuracy of the classification, yet there are practical and industrial needs to correctly predict healthy (negative) thyroid condition data, which are mostly medical examination results, and minimize false-negative rates under the prediction of healthy thyroid conditions.

Methods: The data sets included thyroid and comprehensive medical examination reports. The textual data are not only documented in fully complete sentences but also written in lists of words or phrases. Therefore, we propose static and contextualized ensemble NLP network (SCENT) systems to successfully reflect static and contextual information and handle incomplete sentences. We prepared each convolution neural network (CNN)-, long short-term memory (LSTM)-, and efficiently learning an encoder that classifies token replacements accurately (ELECTRA)-based ensemble model by training or fine-tuning them multiple times. Through comprehensive experiments, we propose 2 versions of ensemble models, SCENT-v1 and SCENT-v2, with the single-architecture–based CNN, LSTM, and ELECTRA ensemble models for the best classification performance and practical use, respectively. SCENT-v1 is an ensemble of CNN and ELECTRA ensemble models, and SCENT-v2 is a hierarchical ensemble of CNN, LSTM, and ELECTRA ensemble models. SCENT-v2 first classifies the 3 labels using an ELECTRA ensemble model and then reclassifies them using an ensemble model of CNN and LSTM if the ELECTRA ensemble model predicted them as "healthy" labels.

Results: SCENT-v1 outperformed all the suggested models, with the highest F1 score (92.56%). SCENT-v2 had the second-highest recall value (94.44%) and the fewest misclassifications for caution-required thyroid condition while maintaining 0 classification error for the critical thyroid condition under the prediction of the healthy thyroid condition.

Conclusions: The proposed SCENT demonstrates good classification performance despite the unique characteristics of the Korean language and problems of data lack and imbalance, especially for the extremely low amount of critical condition data.

The result of SCENT-v1 indicates that different perspectives of static and contextual input token representations can enhance classification performance. SCENT-v2 has a strong impact on the prediction of healthy thyroid conditions.

(JMIR Med Inform 2021;9(9):e30223) doi:10.2196/30223

KEYWORDS

deep learning; natural language processing; word embedding; convolution neural network; long short-term memory; transformer; ensemble; thyroid; electronic medical records

Introduction

In South Korea, a large portion of medical services are maintained and operated under the public health insurance system [1-4], and the Korean National Health Insurance Corporation conducts biannual national health screening examinations. Apart from government-sponsored biannual health examination services, which are different from the health insurance system in the United States, Korean companies provide regular medical checkups to their employees annually according to Article 43 of the Occupational Safety and Health Act [5]. The entrusted companies conduct the examination in partnership with affiliated examination centers in large hospitals or professional examination centers and collect the results from individual medical institutions to provide follow-up health care services to the clients.

Electronic medical records (EMRs) and other forms of medical documentation are designed to focus on the convenience of work for medical personnel in line with the primary use of patient care. The text records of any examination numerical values and comprehensive findings provided by more than 1 examination institution are not standardized and are written in nonunified formats with different periods and health professionals. Thus, to ensure that consistent services are offered, a group of experienced nurses in examination work has been established using classification guidelines based on important keywords and by manually classifying individual test results to organize these results into a single unified format. In this study, thyroid ultrasonography and hormone tests were selected among the various measurements for the application of ensemble language models. The following sections are targeted for this study: individual text diagnosis of thyroid diseases, 3 numeric variables for thyroid hormone examination results, and comprehensive medical examination reports, including doctors' comments.

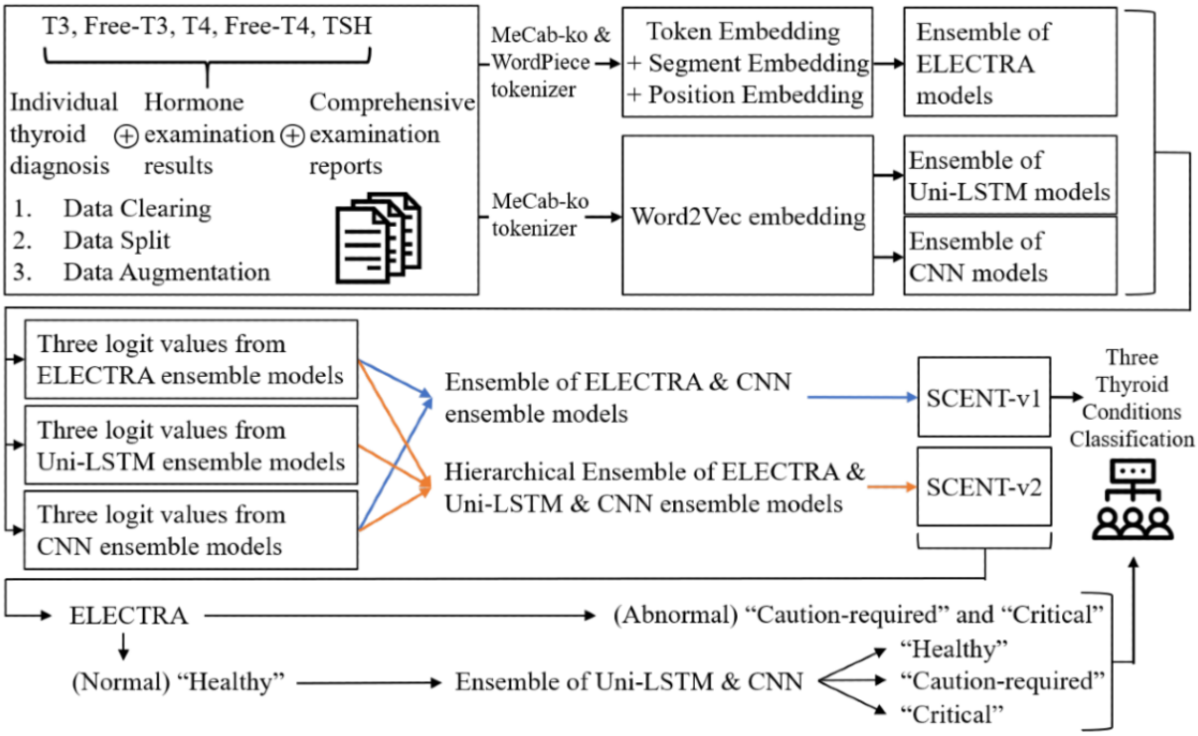
When the rule-based text classification is considered for the analysis of contents in EMRs, repetitive classification and human labor–intensive verification can be required for an extensive rule set, regular expression, and branch logic because of a data model that is not designed for secondary usage of text data or sharing and interworking between multiple agencies [6-8]. However, various implementations in medical natural language processing (NLP) and applications of diverse language models can be considered with recent advances in NLP and techniques based on artificial neural networks [9-16] for data extraction, early detection of diseases, diagnostic support, and prediction of outcomes. Deep learning (DL) models represent intricate structures in large data sets by updating the internal parameters from backpropagation. Such learning techniques produce promising results in various tasks in processing images, videos, audio, and text data [17].

The data sets in our study are textual data that describe the findings and doctors' comments from thyroid ultrasonography and additional comprehensive medical examination results. Such textual data can be considered and processed using NLP methods in DL. Referring to Wu et al [9], the most widely used DL model is recurrent neural network (RNN) variants, while Word2Vec [18] is the most common in embedding architectures. Among their reviewed papers, text classification has the highest percentage (41.5%) for clinical NLP tasks, followed by bidirectional encoder representations from transformers (BERT) [19]. BERT can be used by either training from scratch, directly using fixed pretrained models, or fine-tuning it.

In this study, we initially developed multiple single-architecture–based deep neural network models in NLP not only by using the efficiently learning an encoder that classifies token replacements accurately (ELECTRA) [20] model, which is a pretrained model with open Korean corpus data sets [21] in our study, but also by inventing a convolutional neural network (CNN) [22] model and long short-term memory (LSTM) [23] model. We chose the ELECTRA language model, which has an identical structure to BERT, because it achieves better performance on various NLP benchmarks than BERT and verifies that different pretraining methods are more effective for downstream NLP tasks. However, ELECTRA has a sequence limitation of 512 input tokens; thus, the LSTM structure is employed to capture the full length of contextual representations of input words. For the ELECTRA model, we propose a keyword-based trimming method for the comprehensive medical examination section of the input data sets to reflect thyroid-related information, which could be compulsively truncated because of limitations, effectively for the contextual representations.

Furthermore, we investigate and establish ensemble classification models based on the CNN, LSTM, and ELECTRA models. The combination of static and contextual NLP models is required not only to capture different perspectives of static and contextual word representations from the same input sequences but also to consider the characteristics of the data. The format of the data sets is not standardized or unified; thus, they can be prepared as complete sentences, lists of terminology-based words or phrases with or without numbering them, and groups of numerous medical examination measurements. Such aspects can be an obstacle, particularly for training the contextual relationships between input word tokens. Consequentially, we propose ensemble models to capture static and contextualized input word representations of textual examination data and classify them into 3 labels: healthy, caution required, and critical thyroid conditions. We construct 2 ensemble models and call them static and contextualized ensemble NLP network (SCENT) systems. SCENT version 1, SCENT-v1, is an ensemble or soft voting method for the CNN and ELECTRA ensemble models. SCENT-v2 is a hierarchical ensemble of CNN, LSTM, and ELECTRA ensemble models. SCENT-v2 initially classifies the 3 thyroid conditions using the ELECTRA ensemble model and reclassifies the selected labels, only if the ELECTRA ensemble model predicted them as "healthy" thyroid conditions, using an ensemble of CNN and LSTM ensemble models (Figure 1).

Figure 1. Overall flow of our proposed ensemble approach. T3: triiodothyronine; Free-T3: free triiodothyronine; T4: thyroxine; Free-T4: free thyroxine; TSH: thyroid stimulating hormone; ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately; Uni-LSTM: unidirectional long short-term memory; CNN: convolution neural network; SCENT: Static and Contextualized Ensemble NLP-neTworks; -v1: version 1; -v2: version 2.


T3, Free-T3, T4, Free-T4, TSH
Individual thyroid diagnosis
Hormone examination results
Comprehensive examination reports
1. Data Clearing
2. Data Split
3. Data Augmentation
MeCab-ko & WordPiece tokenizer
MeCab-ko tokenizer
Token Embedding + Segment Embedding + Position Embedding
Word2Vec embedding
Ensemble of ELECTRA models
Ensemble of Uni-LSTM models
Ensemble of CNN models
Three logit values from ELECTRA ensemble models
Three logit values from Uni-LSTM ensemble models
Three logit values from CNN ensemble models
Ensemble of ELECTRA & CNN ensemble models
Hierarchical Ensemble of ELECTRA & Uni-LSTM & CNN ensemble models
SCENT-v1
SCENT-v2
Three Thyroid Conditions Classification
ELECTRA
(Abnormal) "Caution-required" and "Critical"
(Normal) "Healthy"
Ensemble of Uni-LSTM & CNN
"Healthy"
"Caution-required"
"Critical"

Methods

Data Labeling Using Thyroid Ultrasonography Keywords

Thyroid glands are butterfly-shaped endocrine glands located in the lower front of the neck and are responsible for the production of thyroid hormone [24]. Thyroid nodules are lumps produced by abnormal growth of thyroid cells that appear as either solid (hard lumps) or cystic (water lumps). If nodules are found in the thyroid gland during a medical examination, thyroid ultrasonography can be performed to check for signs of cancer. It is also possible to check thyroid hormone levels and conduct blood tests on thyroid antibodies to identify other types of thyroid disorders [25]. Thyroid nodules typically do not cause symptoms or require treatment, but a small number of thyroid nodules can be diagnosed as cancerous. Thyroid cancer is mainly detected and diagnosed using blood tests and thyroid ultrasonography. Thyroid ultrasonography may show the size and shape (solid or liquid-filled cysts) of thyroid nodules.

For our experimental data sets, to minimize classification errors, an experienced nurse with expertise in the field of health examination performed the first labeling task, and a member of another nurse group performed the second labeling of each entry. After that, reclassification proceeded through group discussions on the parts with differences in classification. In this study, the final classification tags for each entry were used as labels. The basic test results classification criteria are defined as follows:

- Healthy: no abnormalities (normal), simple cyst, tubular cyst, thyroid resection (thyroidectomy), benign calcification.
- Caution required: hypothyroidism, unequal parenchyma, internal thyroid disease, thyroiditis, nodule, thyromegaly, hyperechoic lesion, hypoechoic lesion, hyperechoic nodules, hypoechoic nodules, cystic lesions.
- Critical: tumor, malignant, biopsy, fine-needle aspiration cytology.

Data Preprocessing

The data sets, which consist of individual text diagnosis of thyroid diseases, comprehensive medical examination text reports including doctors' comments, and 3 categorical variables for individual hormone examination results, were classified as healthy, caution required, and critical labels in total. The categories of hormone examination results were classified as normal or abnormal by comparing the results of the numerous subtests for triiodothyronine (T3), free triiodothyronine (Free T3), thyroxine (T4), free thyroxine (Free T4), and thyroid-stimulating hormone with the reference range for each device and test. A total of 122,581 textual data were collected in the free form of EMRs from 284 health care institutions in the Republic of Korea between January 2015 and May 2020; thus, data clearing was compulsory. The data sets were written in Korean with numerous English biological and chemical terminologies, including various special characters. Many special characters and measurement units with brackets such as “blood pressure 120/80 mm/Hg”, “microalbuminuria is less than 30 mg/g”, and “renal cyst (left side 1.4 cm)” can increase vocabulary size and lengthen the sequence of input texts unnecessarily. Therefore, Korean, English, numerical characters, and only selected special characters, such as “%”, “'”, “/”, “~”, “^{2}”, “-”, “,”, and “.” remained after preprocessing. In addition, the 3 dummy variables of hormone examination were converted concisely into 3 sentences before tokenization: “hormone examination results were normal,” hormone examination results were abnormal,” and “hormone examination was not conducted.”

Among the total sample size of 122,581 text data, 84,111 samples, 37,220 samples, and 1250 samples were labeled as healthy, caution required, and critical conditions, respectively. The extreme data imbalance can be troublesome for training or fine-tuning the DL models, so the least amount of critical condition data was initially divided into 7:1:2 ratios for training, validation, and test data sets. The training data were then augmented by splitting sentences and each sentence was attached one by one starting from the first sentence to the last. For instance, a sample datum with 3 consecutive sentences was multiplied into 3 samples with the first 1 sentence, the first 2 sentences, and the entire 3 sentences each from the original sample data. During the augmentation, the order of sentences was preserved as the original sample data because split sentences were added in the order of original sequences. Consequently, the critical condition data sets were split and then augmented, and the healthy and caution-required condition data sets were only divided according to the ratio of prepared data (Table 1). The training data sets for the critical condition were augmented from 875 to 29,174 samples. After that, the entire prepared training data sets were randomly shuffled. Relatively short examples of data and translations for each class are listed in Table 2. The data sets consist of a sequential combination of individual diagnosis, hormone examination results, and comprehensive medical examination reports. Comprehensive reports are occasionally omitted.

Table 1. Numbers of divided sample data sets. Only train data for critical thyroid condition are augmented and the original amount of data before the augmentation is given in brackets (N=122,581).

Thyroid conditions	Total number of prepared data sets			
	Train (n=87,524), n (%)	Validation (n=21,119), n (%)	Test (n=42,237), n (%)	Total, n (%)
Healthy	29,175 (33.33)	18,312 (86.71)	36,624 (86.71)	84,111 (68.62)
Caution required	29,175 (33.33)	2682 (12.70)	5363 (12.70)	37,220 (30.36)
Critical	29,174 [875] (33.33)	125 (0.59)	250 (0.59)	1250 (1.02)

Table 2. Short examples and English translations for each thyroid condition.

Examples	Contents
Healthy condition	
Original	정상. 호르몬 검사 수치 정상입니다. uibc 감소, 철 증가, 총 콜레스테롤 증가, glucose증가, 골다공증.
Translation	Normal. Hormone examination results were normal. UIBC decreases, iron increases, total cholesterol increases, glucose increases, osteoporosis.
Original	정상. 호르몬 검사 수치 미 판정입니다. 체중 관리에 주의 가 필요합니다. 총 콜레스테롤 수치가 높습니다. 중성지방수치가 높습니다. 저밀도 콜레스테롤 수치 가 높습니다 .
Translation	Normal. Hormone examination was not conducted. Please be aware of weight management. Total cholesterol level is high. Neutral fat level is high. Low-density lipoprotein cholesterol level is high.
Caution-required condition	
Original	갑상선염. 호르몬 검사 수치 정상입니다. b형 간염 항체 미 형성. 갑상선염. 고 음영 유방, 유방 양성 석회화 양측.
Translation	Thyroiditis. Hormone examination results were normal. Hepatitis B antibody not formed. Thyroiditis. Dense breast, positive calcification for both.
Original	갑상선염 의심 또는 치유 반흔. 호르몬 검사 수치 정상입니다. 양측 치밀 유방 2. 갑상선염 의심 또는 치유 반흔 3. 담낭 결석 및 콜레스테롤 용종 4. 위염 5. 자궁경부 염 6. a형간염 항체 없음.
Translation	Suspect thyroiditis or scars. Hormone examination results were normal. Dense breasts for both. 2. Suspect thyroiditis or scars 3. Gallstone and cholesterol polyps 4. Gastritis 5. Cervicitis 6. No antibody for hepatitis A.
Critical condition	
Original	갑상선 초음파 검사상 좌엽 결절 2.78 cm 소견입니다. 세침 흡인 세포검사를 받으시 길 권유합니다. 호르몬 검사 수치 미 판정입니다.
Translation	Thyroid ultrasonography shows 2.78 cm of left nodule. We recommend taking a fine needle aspiration cytology. Hormone examination was not conducted.
Original	갑상선 좌측부에 10.2mm 크기의 저 에코결절이 1개 있으며 감별 진단을 위해 세침검사로 확인 요망됨. 결론은 좌측 부 갑상선 결절. 요망 세침검사로 확인 및 의사와 상담 요망. 호르몬 검사 수치 정상입니다. 위장 조영촬영결과 유 소견입니다. 갑상선 초음파 검사 결과 유 소견입니다.
Translation	There is 10.2mm size of 1 hypoechoic nodule in left-sided thyroid and requires fine needle aspiration cytology for differential diagnosis. Left-sided thyroid nodule in the conclusion. Have consultations with doctors and confirm with fine needle aspiration cytology. Hormone examination results were normal. Blood sugar level before a meal is high. Upper gastrointestinography results were abnormal. Thyroid ultrasonography results were abnormal.

Tokenization

Korean is an agglutinative language and one of the morphologically rich [26] and typologically diverse [27] languages; a character is composed of consonants and vowels of the Korean alphabets in 3 positional forms: choseong (syllable onset), jungseong (syllable nucleus), and jongseong (syllable coda). The positional forms are displayed in the lexicographic order of Korean alphabets as follows:

Choseong: ㄱ ㄲ ㄴ ㄷ ㄸ ㄹ ㅁ ㅂ ㅃ ㅅ ㅆ ㅇ ㅈ ㅉ ㅊ ㅋ ㅌ ㅍ ㅎ

Jungseong: ㅏ ㅐ ㅑ ㅒ ㅓ ㅔ ㅕ ㅖ ㅗㅘㅙㅚㅛㅜㅝㅞㅟㅠㅡㅢ ㅣ

Jongseong: (None)ㄱㄲㄳㄴㄵㄶㄷ ㄹㄺㄻㄼㄽㄾㄿㅀㅁ ㅂ ㅄㅅㅆㅇㅈㅊㅋㅌㅍㅎ

One of the common challenges in text preprocessing for Koreans is the ambiguity of word spacing, unlike other languages. For example, an English phrase “Be able to do” is translated into a grammatically accurate Korean phrase “할 수 있다,” which has 2-word spaces. When not strictly aware of Korean orthography, it can also be written as “할수 있다” (Beable todo) with 1-word space or “할수있다” (Beabletodo) without any word space. Furthermore, various postpositions or particles, which means “helping words” in English, are immediately attached after nouns or pronouns without any white space. For instance, English phrases “I am” and “You and me” become “Iam” and “Youand me” in Korean phrases. This can make it difficult to decompose sentences into distinguishable morphemes; for example, the same noun(s) or pronoun(s) can be tokenized into multiple tokens, even if their actual meaning may not differ. Such inconsistent grammatical errors and unique grammatical aspects can cause the same expression of word-level texts to be tokenized into different tokens, which may result in difficulty in training NLP models.

To resolve such problems, we used the MeCab-ko [28] tokenizer, which was originally introduced as MeCab for Japanese morphological analysis by Kudo et al [29]. The variation for the Korean tokenizer yields good performance to handle such problems by reconstructing and unifying a grammatical structure with a relatively faster speed than other Korean tokenizers [30]. WordPiece [31-33], which was originally introduced for Japanese/Korean segmentation, was employed for the transformer [34] encoder–based models such as BERT and ELECTRA for various purposes. One of the major advantages is that it can increase the robustness against the out-of-vocabulary (OOV) problem with a relatively small vocabulary size by disassembling words into subword units using a given text corpus. Therefore, in this study, we used the combination of MeCab-ko and WordPiece to pretrain the thyroid text data sets to fine-tune the ELECTRA [21] model, which was pretrained with a Korean open-source corpus. The input sentences were initially identified and reconstructed into possible grammatical morphemes by MeCab-ko and then segmented into divisible subwords that maximize the log likelihood of a language model by WordPiece. For instance, "임상적으로," which means "clinically" in English, can eventually be tokenized into "임상," "##적," "으로" as "clinic," "##al," and "ly," where the last part was initially separated by MeCab-ko and the first and second parts were segmented by WordPiece after that.

The average and maximum lengths of the input sequence resulting from different tokenizers are listed in Table 3. All input tokens from every sample had right-skewed (positive skewness) distributions. To reduce the sequence length, especially for the ELECTRA model, which has a limitation of 512 tokens, only comprehensive examination reports were trimmed for every sample. Based on the first sentence containing the word "thyroid," all subsequent sentences including 1 previous sentence were extracted and then recombined with individual text diagnosis of thyroid diseases and textualized hormone examination results. Original comprehensive text reports were used when the word "thyroid" does not exist.

Table 3. Comparison of different tokenizers and the numbers of input tokens.

Tokenizer	Average number of tokens			Maximum number of tokens		
	Train	Valid	Test	Train	Valid	Test
MeCab-ko	494.2	522.3	520.1	2227	2219	2240
WordPiece for BERT[a]	664.7	698.7	695.9	4096	2943	4171
WordPiece for ELECTRA[b]	564.9	596.3	593.7	2656	2472	2500
MeCab-ko and WordPiece	540.6	570.0	567.6	2608	2431	2435
MeCab-ko (trimmed[c])	370.4	419.6	418.8	2162	2219	2216
MeCab-ko and WordPiece (trimmed[c])	404.9	457.6	456.6	2365	2431	2412

[a] BERT: bidirectional encoder representations from transformers.

[b] ELECTRA: efficiently learning an encoder that classifies token replacements accurately.

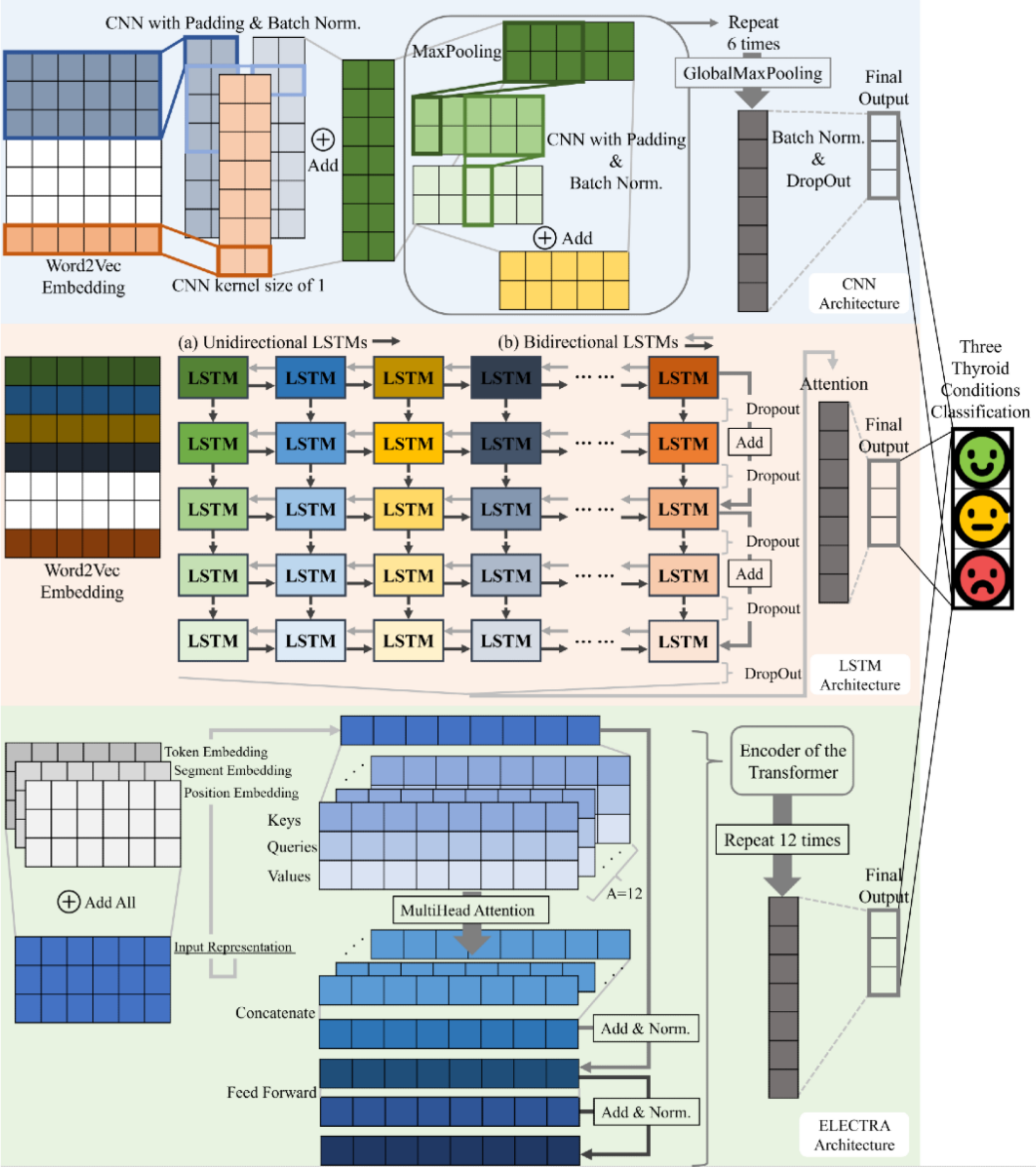
[c] Trimmed: The data sets were trimmed based on the keyword "thyroid" in the comprehensive medical examination text part.

Proposed Framework

Overview

In this study, we propose ensemble models, SCENT-v1 and SCENT-v2, which can reduce generalization errors of the prediction and reflect static and contextual perspectives of word representations in accordance with thyroid and general examination reports. Our proposed ensemble models consist of multiple single-architecture–based ensemble models from CNN, LSTM, and transformer encoder architectures as shown in Figures 1 and 2. We initially created a CNN with batch normalization (BN) [35] transform approach, LSTM with 2 shortcut connections [36] including an attention mechanism [37], and ELECTRA models. Each model was trained or fine-tuned 10 times with different settings of epochs, learning rates, and batch sizes. Subsequently, each single-architecture–based model was trained or fine-tuned 10 times and then combined into an ensemble model. In other words, 3 respective CNN, LSTM, and ELECTRA ensemble models were constructed by combining each of the 10 model's prediction averages from softmax functions, or simply by performing soft voting, to stabilize the variances of classification performance. Based on the experimental results of single-architecture–based ensemble models, we selected the CNN-Word2Vec, Uni-LSTM, and ELECTRA-v2 ensemble models for further ensemble approaches. The 3 distinct models were CNN with trainable Word2Vec embedding, unidirectional LSTM with trainable Word2Vec embedding, and the second version of ELECTRA fine-tuned with trimmed data sets.

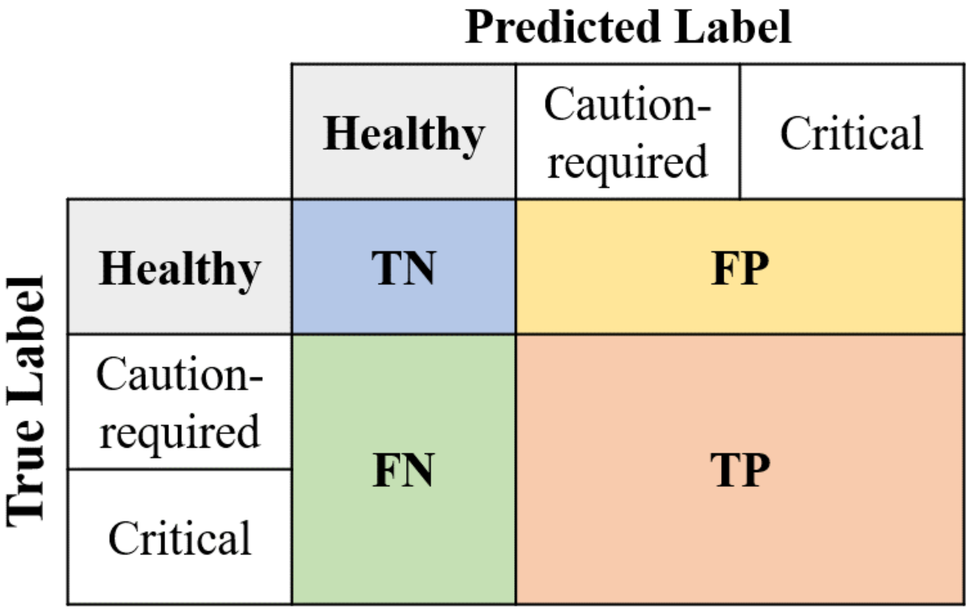
Figure 2. The architecture of the proposed ensemble models. Each model is trained or fine-tuned ten times for each ensemble model. Best viewed in color. CNN: convolution neural network; Batch Norm.: batch normalization transforms; LSTM: long short-term memory; ELECTRA: Efficiently Learning an Encoder that Classifies Token Replacements Accurately; Norm.: layer normalization.



The final predictions for the thyroid condition classification were then determined using ensemble and hierarchical ensemble methods, namely, SCENT-v1 and SCENT-v2, respectively. In this experiment, static word representations were captured from the CNN-Word2Vec ensemble model, and contextualized word representations were captured from the Uni-LSTM ensemble model with the ELECTRA-v2 ensemble model, which exclusively considers the initial 512 token sequences in the trimmed data sets. SCENT-v1 is an ensemble of CNN and ELECTRA ensemble models, and SCENT-v2 is a hierarchical ensemble of CNN-Word2Vec, Uni-LSTM, and ELECTRA-v2 ensemble models (Figure 1). The multilabel classification in SCENT-v2 was based on the 3 thyroid condition predictions from the ELECTRA-v2 ensemble model and reclassified the selected labels using an ensemble of CNN-Word2Vec and Uni-LSTM ensemble models, where only the ELECTRA-v2 ensemble model predicted "healthy" thyroid conditions. In other words, SCENT-v2 kept the decisions from the ELECTRA-v2 ensemble model for "caution required" and "critical" thyroid condition predictions, and then made final decisions from an ensemble of CNN-Word2Vec and Uni-LSTM ensemble models only for the "healthy" thyroid conditions, which were predicted by the ELECTRA-v2 ensemble model.

Our proposed SCENT-v2 is designed for the industrial purpose in that it saves time and cost by reducing the number of manual thyroid condition classification steps required and human misclassification errors. Perfect overall classification accuracy for current and future data sets must be the ideal solution. However, there are numerous obstacles such as imbalanced numbers of data sets and the difficulty level of the problem. This hierarchical ensemble method, therefore, was pursued to minimize the numbers of false negatives and maximize the numbers of true negatives as depicted in Figure 3. It primarily aimed to correctly predict an exceedingly high number of healthy thyroid conditions with 100% precision of healthy (negative) thyroid labels and leave the remaining data sets for manual classification to provide precise health care services and reduce the human classification workloads. This approach was proposed to take into account aspects of practical and industrial usage efficiency by sacrificing the overall accuracy but reducing the large manual workloads. Based on the validation data sets, among the 3 single-architecture–based ensemble models, the ELECTRA-v2 ensemble model indicates a relatively low number of false positives, and the 2 CNN-Word2Vec and Uni-LSTM ensemble models show a relatively small number of false negatives in the prediction of healthy thyroid conditions (Figure 3). Accordingly, we constructed SCENT-v2 for the hierarchical ensemble model in this study.

Figure 3. A confusion matrix for healthy thyroid condition datasets. TN: true negative; FP: false-positive; FN: false-negative; TP: true positive.



		Predicted Label		
		Healthy	Caution-required	Critical
True Label	Healthy	TN	FP	
	Caution-required	FN	TP	
	Critical			

Embedding

Word embedding is a way of expressing words that are converted into distributed vector representations. Mikolov et al [18] introduced Word2Vec embedding, which provides remarkable performance for capturing syntactic and semantic word relationships. Continuous bag of words (CBOW) and Skip-gram methods were proposed with several loss function approaches in their paper, and we used skip-gram with negative sampling (SGNS) Word2Vec in our NLP models. For a given corpus sequence T length of words $w_1, \ldots, w_{t-1}, w_t, w_{t+1}, \ldots, w_T$, where the training context size is c, CBOW predicts the probability of the current word w_t as $P(w_t|w_{t-c}, \ldots, w_{t+c})$. By contrast, the skip-gram method predicts the probability of the context words as $P(w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}|w_t)$ by the softmax function calculated as

where v_w and $v\ _w$ are the input and output word vector representations, respectively; $|V|$ is the vocabulary size; and w_O and w_I refer to the target word representations and the given word representations, respectively. Negative sampling is suggested as an alternative to the initially used hierarchical softmax function because of the cost of computing the vocabulary size. It is defined by the objective function calculated as

where every $\log P(w_O|w_I)$ in the objective is replaced. The probability $P_n(w_i)=f(w_i)^{3/4}/\sum_{j=0}[f(w_j)^{3/4}]$ is a unigram distribution that allows the use of a selected number of n negative samples instead of the number of vocabulary sizes. To use the SGNS Word2Vec, we initially predefined 5 context sizes, 5 negative samples, and 300 dimensions for each vector representation. The embedding was then pretrained unsupervised using Wikipedia corpus data [38], which contain 162,861 articles on various topics. The grammatical expressions in the corpus data were restructured and prepared using the MeCab-ko tokenizer before pretraining. This method helps convert the ith word w_i to a fixed length of 300-dimensional word vector x_i, and thus, can be calculated algebraically; for instance, *vector(“한국$_{Korea}$”)–vector(“서울$_{Seoul}$”)+vector(“도쿄$_{Tokyo}$”)* results in a vector representation with most similarity of the word 일본$_{Japan}$ (the subscripts are English translation).

Word embeddings for transformer-based models BERT and ELECTRA have a different approach for establishing word vocabulary because of the tokenizer called WordPiece. Rather than the n-gram strategy in Word2Vec, this approach initializes the vocabulary with its size to include all character representations in each corpus by using a greedy longest-match-first [39] approximation, which picks the longest subwords or prefixes inside the corpus. It selects a new word piece that maximizes the log likelihood for the corpus when the word piece is added to the language model. For example, a word piece “un” is added to the vocabulary if the probability of “un” divided by “u” and “n” is higher than other subword units. After the preparation of token embeddings, both transformer-based models create their input representations by summing up the token, segment, and position embeddings. Two special tokens were used to distinguish sentences. A special classification token (CLS) was inserted as the first token of all sentences for the classification task, and a special separator token (SEP) was used to distinguish sentence pairs as the first and second sentences, where the segment embedding distinguishes them. The position embedding shows the location of each token as $PE_{(p,2i)}=\sin(p/10000^{2i/d})$ and $PE_{(p,2i+1)}=\cos(p/10000^{2i/d})$, where p indicates the location of the embedding vector in the input sentence; and i is the index of the dimension within the embedding vector. A hyperparameter d=768 indicates the dimensions of all corresponding embedding vectors and encoder layers of the transformer.

Convolution Neural Networks

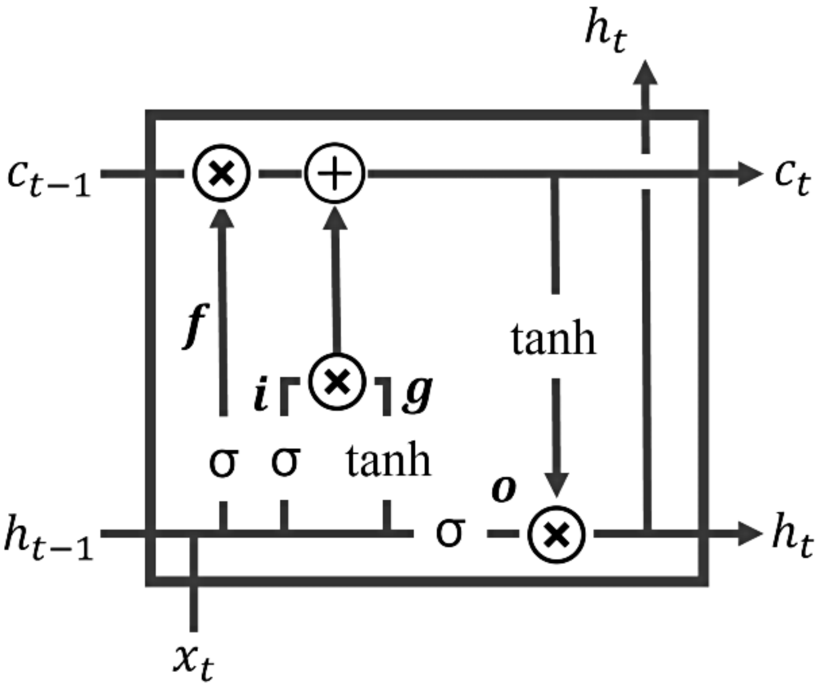
CNN can be described as a structure that is originally designed for processing images to identify patterns of features by weight sharing and local connectivity. CNN can be used for NLP as well and extracts the same features regardless of positions by sliding CNN filters over consecutive tokens with a fixed window size. CNNs have become an essential method in computer vision tasks [40-42] and produce good results on sentence classification tasks [43]. In this study, we suggest deep CNN feature–learning methods to determine how static word vector representations are achieved in text classification. The model, which is depicted at the top of Figure 2, considers input word tokens through pretrained Word2Vec, where the maximum length of input sequences is set to 2240 tokens. The CNN model initially vectorizes input word tokens through word embeddings with a dimensionality of 300 for each vector representation. A convolution operation then generates a feature map $c=f(Wx+b)$, where W and b are the weight and bias parameters of the model, respectively, and $f(\cdot)$ is a nonlinear function such as rectified linear units [44] and $\text{ReLU}(x)=\max(0, x)$. We employed the BN transform in the convolutional operation before the nonlinearity function. In this study, we used the BN transform in the CNN operations because it [35] can reduce the necessity for dropout [45], and other methods such as L_2 regularization become ineffective when combined with BN, but only influence learning rates [46].

Starting from the lower layers of the CNN model, we conducted the summation of 2 consecutive 3 kernel sizes of convolution layers with BN and 1 kernel size of the convolution layer without BN from pretrained SGNS Word2Vec. The word vectors with a dimensionality of 300 are represented as local features of word vectors with 250 dimensions. The structure then connects to a max-pooling combination consisting of size 3 and stride 2 of max-pooling, 2 consecutive 3 kernel sizes of convolution layers with BN, and a simple shortcut connection with a consistency of 250 dimensionality. The combination was repeated 6 times to determine deep representations of static word features, and a global max-pooling operation extracted the maximum values over the dimensions. The penultimate layer was then connected to the softmax computation layer for the label prediction using BN with a dropout rate of 0.5. The CNN model was constructed with 3 variants of word embedding: CNN-random, CNN-fixed-Word2Vec, and CNN-Word2Vec. The only difference is that the parameters of the embedding part were randomly initialized, transferred from pretrained SGNS Word2Vec, maintained nontrainable, and fine-tuned pretrained SGNS Word2Vec during model training.

Long Short-term Memory

RNN can be described as a neural network that learns from sequential data such as time-series data. It has a recurrent structure that learns temporal or sequential patterns and makes the information persistent. However, gradient vanishing is a significant problem while training RNN-based models, and it can cause a long-range dependency when a long input sequence is given. LSTM is a form of RNN structure with added gates in the LSTM interface (Figure 4). Memory cell block alleviates long-term dependency problems. In a unit of LSTM, the forget gate $f_t=\sigma[W_f(h_{t-1}, x_t)+b_f]$ decides how much to neglect when the previous hidden state and the vector x_t at time t are given. The new memory node $g_t=\tanh[W_g(h_{t-1}, x_t)+b_g]$ stores new information from the previous hidden state and the vector x_t. The input gate $i_t=\sigma[W_i(h_{t-1}, x_t)+b_i]$ decides how much new information can be accommodated by element-wise multiplication of the new memory cell. The output gate $o_t=\sigma[W_o(h_{t-1}, x_t)+b_o]$ determines how much information is delivered to the hidden state h_t at time t. In conclusion, the hidden state $h_t=o_t\cdot\tanh(c_t)$ is produced with the element-wise multiplication of the output gate and memory cell of c_t, where the memory cell $c_t=f_t\cdot c_{t-1}+i_t\cdot g_t$ is produced by the summation of the element-wise multiplication of the forget gate with previous memory cell of c_{t-1} and element-wise multiplication of the input gate with the new memory node. $\sigma(x)=1/(1+e^{-x})$ is a sigmoid function, $\tanh(x)=(e^{2x}-1)/(e^{2x}+1)$ is a hyperbolic tangent function, and W and b are distinguishable weight and bias parameters, respectively.

Figure 4. One sample unit of long short-term memory. x: vector; h: hidden state; f: forget gate; i: input gate; g: memory node; o: output gate; c: memory cell; σ: sigmoid function; tanh: hyperbolic tangent function.



As shown in the middle of Figure 2, for the contextual LSTM model, we constructed 5 unidirectional LSTM layers with 650 units per layer with 2 shortcut connections and a 50% dropout rate on the nonrecurrent connections for every LSTM layer. The bidirectional LSTM model follows the same structure, but each forward and reverse LSTM has 310 units, and 620 units are concatenated per layer. The 2 shortcut connections, which can help prevent the models from overfitting, are linked from the first to the third LSTM layers and from the third to the fifth LSTM layers. We then apply an attention mechanism that can measure the importance of the given tokens before thyroid classification. A hidden representation u_i=tanh($W_u \cdot h_i + b_u$) from the last hidden layers of LSTM is calculated, and a weighted summation vector $v=\sum_i \alpha_i h_i$ is determined by attention as follows:

where W_u and b_u are the weight and bias parameters, respectively; and u_c is a context vector that is randomly initialized and jointly learned. The weighted vector then passes to the last layer of this model to compute the softmax probabilities of each thyroid condition. Both Uni-LSTM and Bi-LSTM models vectorize input tokens using the MeCab-ko tokenizer and use trainable pretrained SGNS Word2Vec embedding.

Transformer

RNN-based models take a long time to compute input sentences because the calculations are performed sequentially. However, transformer processes input sentences in parallel and capture various relationships between words in a sentence with the help of a multihead self-attention mechanism. Because the input tokens are not computed sequentially, transformer includes special position embedding that reflects position information in the attention mechanism to construct word-to-word importance and dependency. The BERT and ELECTRA models are based on the transformer. The authors of the transformer proposed the architecture of encoder and decoder with a unique attention mechanism. Both BERT and ELECTRA, which are pretrained BERT and ELECTRA, respectively, in our study, use multiple encoder layers of the transformer exclusively, as shown at the bottom of Figure 2. After summing up the 3 tokens, segments, and position embeddings as described above, the transformer encoder obtains linear projections of key, query, and value for each input representation. The scaled dot-product attention is calculated through

Attention scores are obtained from each query projection by keys, attention weight distribution is computed through a softmax function, and the final values are obtained through the product of the value projection. This attention step is repeated A=12 times and concatenated to Concat(head$_1$, ..., head$_{12}$)W^O from head$_a$=attention(KW_a^K, QW_a^Q, VW_a^V), where the dimensions are d_k=d_q=d_v=64, and the distinguishable weights are W^O, W_a^K, W_a^Q, and W_a^V. This can help train the model in which the same input tokens can be represented from multiple perspectives. The results from multihead attention are then connected to 2 layers of feed-forward neural networks FFNN=ReLU(0, xW_1+b_1)W_2+b$_2$, where the shape of W_1 is (d=768, d_{ff}=3072) and W_2 is (d_{ff}=3072, d=768), and processed with residual connection and layer normalization, as depicted with arrows in Figure 2. In conclusion, these encoder layers are stacked 12 times and then connected to the penultimate layer, which is a dense layer with 768 units. Then, the softmax probabilities are computed for predicting the 3 thyroid conditions. The 2 transformer encoder-based models are pretrained with different learning methods: random masking procedures [19] for BERT and replaced token detection [20] for ELECTRA.

Experimental Settings

The hyperparameters of the different NLP models are listed in Table 4. CNN- and LSTM-based models were trained for 30 epochs with adaptive learning rates by monitoring validation loss; the learning rate decays by a factor of 0.7 if the validation loss is not improved (decreased) within 1 epoch. Transformer encoder–based models, which were initially pretrained with the open-source Korean corpus data, were fine-tuned for 15 epochs with a fixed learning rate. Adam [47] optimizer, where β_1=0.9, β_2=0.999, and ε=1e–8, is considered for all NLP models. The experiments were implemented with TensorFlow [48], PyTorch [49], and Hugging Face [50] libraries, and a GeForce RTX 2080 Ti 11-GB graphic processor unit.

Table 4. Detailed information about different NLP models.

Models	Tokenizer	Embedding vocabulary size	Number of parameters	Initial learning rate	Batch size
Convolution neural network	MeCab-ko	100,000	32 million	1e–3	64
Unidirectional long short-term memory	MeCab-ko	100,000	46 million	2e–4	32
Bidirectional long short-term memory	MeCab-ko	100,000	40 million	2e–4	32
Bidirectional encoder representations from transformers	WordPiece	8002	92 million	2e–5	8
ELECTRA[a]-version 1	WordPiece	32,200	110 million	2e–5	8
ELECTRA-version 2	MeCab-ko & WordPiece	35,000	112 million	2e–5	8

[a] ELECTRA: efficiently learning an encoder that classifies token replacements accurately.

Results

According to Table 5 and Figure 5, the macroaveraged precision, recall, and F1 scores are calculated due to the imbalance of multilabel data sets and confusion matrices, respectively. For the single-architecture–based ensemble models, in general, we observed that CNN-Word2Vec achieved the highest F1 score among the ensemble models, and Uni-LSTM outperformed Bi-LSTM by achieving slightly higher F1 scores. Performance degradation was observed in the CNN-Word2Vec and Uni-LSTM models while training with trimmed data sets, but improvement was observed in ELECTRA-v2. The LSTM architecture has the characteristics of an RNN, and it has connections between units along a temporal sequence. Thus, we assume that there must be a difficulty in learning contextual representations owing to the inconsistency of the data structure: lists of words or phrases and full complete sentences. Although both BERT and ELECTRA models have recorded state-of-the-art results on multiple NLP benchmarks, it is surprising that fine-tuned transformer encoder layer–based models do not achieve the highest F1 score in this classification task even with the highest number of parameters. This is likely because there must be an information loss by input sequence truncation even after the keyword-based trimming method or quality issues about these data sets themselves and the data clearing part.

Table 5. Experimental results from different NLP models. The test results are macroaverage classification values.

Methods (model name) and models	Precision (%)	Recall (%)	F1 score (%)
Convolution neural network (CNN)			
CNN-random[a]	89.33	90.67	89.91
CNN-fixed-Word2Vec[b]	88.01	93.12	90.43
CNN-Word2Vec[c]	92.01	92.87	92.33
Long short-term memory			
Unidirectional long short-term memory	87.23	93.89	90.32
Bidirectional long short-term memory	87.97	92.48	90.09
Transformer encoder			
Bidirectional encoder representations from transformers	86.44	89.69	87.99
ELECTRA[d]-version 1	87.73	92.12	89.82
ELECTRA-version 2	91.03	92.33	91.60
Data trimming			
CNN-Word2Vec (trimmed[e])	90.59	93.56	91.98
Unidirectional long short-term memory (trimmed)	84.77	93.30	88.61
ELECTRA-v2 (trimmed)	89.63	94.47	91.92
Ensemble combination			
CNN-Word2Vec + Uni-LSTM	89.53	94.24	91.76
SCENT[f]-v1: CNN-Word2Vec + ELECTRA-v2 (trimmed)	91.10	94.18	92.56
Unidirectional long short-term memory + ELECTRA-v2 (trimmed)	89.53	94.24	91.76
CNN-Word2Vec + unidirectional long short-term memory + ELECTRA-v2 (trimmed)	91.02	94.19	92.52
Hierarchical ensemble			
CNN-Word2Vec and unidirectional long short-term memory + ELECTRA-v2 (trimmed)	91.30	92.86	91.92
Unidirectional long short-term memory and CNN-Word2Vec + ELECTRA-v2 (trimmed)	86.83	93.88	90.09
SCENT-v2: ELECTRA-v2 (trimmed) and CNN-Word2Vec + unidirectional long short-term memory	89.04	94.44	91.58

[a] Random: randomly initialized embedding.

[b] Fixed-Word2Vec: nontrainable pretrained Word2Vec embedding.

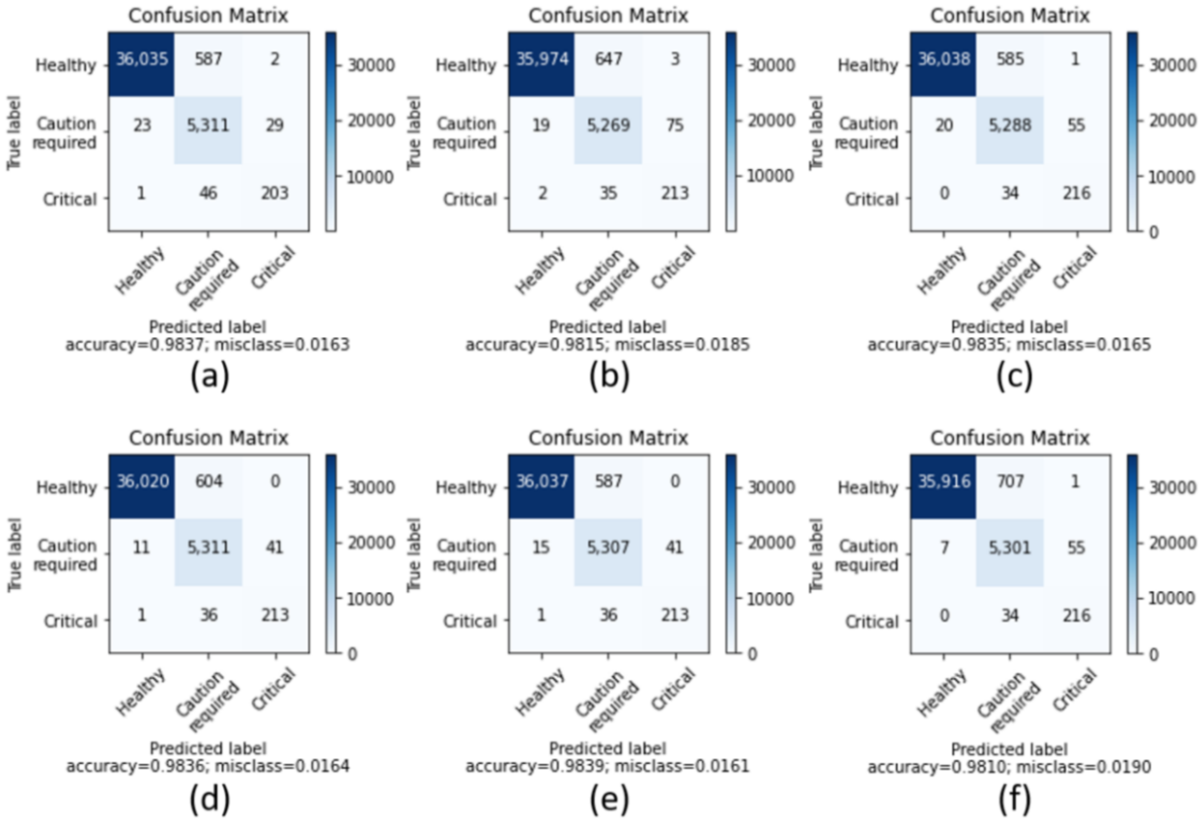
[c] Word2Vec: trainable pretrained Word2Vec embedding.

[d] ELECTRA: efficiently learning an encoder that classifies token replacements accurately.

[e] Trimmed: data sets are trimmed based on the keyword “thyroid” in the comprehensive medical examination text part.

[f] SCENT: static and contextualized ensemble NLP network.

Figure 5. Confusion matrices of multi-label thyroid classification results from the test datasets. All single-architecture-based models are trained or fine-tuned to each ensemble model. The models are (a) CNN-Word2Vec (b) Uni-LSTM (c) ELECTRA-v2 with trimmed data (d) CNN-Word2Vec + Uni-LSTM + ELECTRA-v2 with trimmed data (e) SCENT-v1 (f) SCENT-v2.



SCENT-v1 shows the best performance by calculating the average softmax values, or simply soft voting, from the unnormalized prediction logits of the 2 ensemble models among the NLP models. SCENT-v1 results in 0 misclassifications of healthy thyroid conditions under the prediction of critical thyroid conditions. SCENT-v2 substantially reduced the number of misclassifications of caution-required thyroid condition to the minimum under the prediction of healthy thyroid condition while maintaining 0 misclassifications of critical thyroid condition. According to Figure 5, SCENT-v2 records the highest precision value for the "healthy" thyroid condition among all models, including hierarchical ensemble models. In "Hierarchical Ensemble" section of Table 5, the word "and" distinguishes the base model and the combined model. The base model initially classifies the 3 labels and the other combined model reclassifies selected labels where only the base model is predicted as having "healthy" labels.

The classification results based on tokenizing Korean input sequences into subwords with or without morphological analysis by MeCab-ko differ as represented in the transformer encoder section by the variants of ELECTRA. It may be argued that the number of vocabulary sizes is different in ELECTRA-v1 and -v2; however, the WordPiece tokenizer has a strong effect on OOV, and approximately a 2% increase in F1 score is worthy of close attention. The parameters of word embedding are randomly initialized and pretrained from CNN-random and CNN-Word2Vec, and there are increases in the macroaveraged precision, recall, and F1 scores observed from CNN-random to CNN-Word2Vec. This verifies that transfer learning from a pretrained architecture is an effective and convincing technique for developing deep neural network models. Unlike the validation results in which the false negatives for the healthy thyroid condition (Figure 3) are relatively lower in CNN-Word2Vec and Uni-LSTM ensemble models, the numbers of false negatives from the CNN-Word2Vec, Uni-LSTM, and ELECTRA-v2 (trimmed) ensemble models in the test data sets do not differ. The false positives from the ELECTRA-v2 (trimmed) ensemble model were still lower than those from the other ensemble models. Overall, all ensemble models, including SCENT-v1 and SCENT-v2, showed poor performance in classifying healthy thyroid conditions under the prediction of the caution-required thyroid condition data sets.

Discussion

Limitations

The experiments were originally intended to use only the medical results of the individual thyroid diagnoses. However, the full results of individual text diagnosis of thyroid diseases with hormone examination results and comprehensive medical examination text reports, including doctors' comments, simple body checkups, health care–related guides, and so on, are used as inputs of the models to reduce human curation as much as possible. If the results are labeled as healthy, the keyword "normal" may be mentioned in the reports. In some cases, the results of the examination, which are supposed to be classified as caution required, are labeled as healthy based on the phrase "no change" compared with reports of previous years (1 or 2 years). This can be one of the reasons as to why the number of misclassifications does not dramatically decrease in every experimental model. Furthermore, it cannot guarantee that data clearing was perfectly conducted over the entire nonstandardized 122,581 data sets from 284 health care institutions. It is highly expected that systematic improvement of data quality may enhance all models' performance.

The amount of information in each data varies, and individual or comprehensive finding reports cannot be directly used as a single unit during manual classification. Accordingly, the final decisions were concluded by considering all the data sets. The comprehensive text reports may contain information about thyroid tests regardless of the flow in context, and some are typed manually on a case-by-case basis or automatically filled by enumerating predefined text phrases or sentences depending on the institutions and medical professionals, such as sample data of healthy and caution-required thyroid conditions in Table 2. Depending on the experts, the selection and order of predefined texts may differ for the same thyroid diagnosis. This is partly considered advantageous in deciding thyroid classification by only considering numerous static word representations rather than full contextual word representations and their relationships based on such fragmentary compositions of keywords or phrases in the data sets. This can be a reason as why the single-architecture–based CNN ensemble model achieves the highest F1 score compared with other single-architecture ensemble models of Uni-LSTM and ELECTRA-v2 with trimmed data. However, both contextual models recorded higher recall scores than the static model.

Trimming sentences based on the keyword "thyroid" in comprehensive examination reports because of the limitation of 512 tokens shows an improvement in recall and F1-scores in the ELECTRA-v2 ensemble model. This simple preprocessing, however, cannot guarantee whether the optimal data corresponding to thyroid ultrasonography are used as inputs. We find that the improvement in ELECTRA-v2 indicates that preparing a more suitable data set is meaningful under the sequence length limitations. It is highly expected that the performance of the ELECTRA ensemble model can be further enhanced if the limitation is addressed, and the thyroid ultrasound–related contents can be accurately summarized from comprehensive examination reports. However, performance degradation was observed in the CNN-Word2Vec and Uni-LSTM ensemble models when the same trimming procedure was conducted. This proves that other examination reports in addition to thyroid ultrasound data may have valuable information that can help in the classification of thyroid conditions. This allows us to assume that the decline in health conditions caused by thyroid disease can have an effect related to a person's physical and biological vitality.

Conclusions and Future Research

Our SCENT models show meaningful results despite the lack of data, especially for the critical condition and unique characteristics of Korean, such as auxiliary, adverbial case markers, and word spacing inconsistency. Additionally, our ensemble model methodologies can be applied to data sets with diverse languages and different sequence lengths if only the WordPiece tokenizer is used. Our SCENT models can not only automate the classification of large-scale text data sets at a high speed while maintaining multiclassification performance, but also reduce the human labor force. For SCENT-v1, misclassifying the "critical" case as "caution required" is much less damaging than misclassifying it as "healthy" in this study. However, this model cannot be directly adopted in real-life applications because both type 1 and 2 errors must be considered. Specifically, the false-positive errors under the prediction of caution-required thyroid conditions are too high to be used.

To consider SCENT models for practical use, we preferentially aim to correctly predict the healthy condition labels, which constitute the largest portion among the 3-class data sets. The model SCENT-v2, which is a hierarchical ensemble of CNN-Word2Vec, Uni-LSTM, and ELECTRA-v2 with trimmed data ensemble models, can reduce the number of incorrect classifications of caution-required condition data to a minimum compared with other approaches, while maintaining the number of misclassified critical condition data set to 0 under the healthy thyroid condition prediction. For further studies, the receiver operating characteristic (ROC) and area under the curve (AUC) algorithms, or simply the AUC–ROC curve, can be considered. For the healthy (negative) thyroid classification, the best or optimal threshold value for the classifier based on rest (positive) conditions can be calculated for suitable healthy thyroid prediction performance. Furthermore, as discussed above, the keyword-based trimming method shows that incorporating additional medical results, which are relevant to disease diagnosis and other physical examinations, may enable us to build classification models to outperform the current models that consider only selected examination results: individual text diagnosis of thyroid diseases, hormone examination results, and comprehensive medical examination text reports, including doctors' comments. We may also consider developing DL models that can reflect the results derived from the existing interdisease correlation study [51-53] or causality study [54-57].

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) (Grant Number NRF-2020R1F1A1071527), funded by the Korean government (MSIT; Ministry of Science and ICT). This research was also supported by the Yonsei University Research Fund of 2020 (Grant Number 2020-22-0523). The funding sources played no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Yu-Kyoung Lee (RN-BSN) of AIMMED for data preparation and misclassification verification for this study.

Authors' Contributions

HJK conceived the study and oversaw the technical details regarding electronic medical records. DS and HYK contributed to the architectural design and analysis. M-SJ contributed to data collection, inspection, and preparation. DS prepared the manuscript draft and implemented data preprocessing and experiments. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

References

1. Jeong H. Korea's National Health Insurance--lessons from the past three decades. Health Aff (Millwood) 2011 Jan;30(1):136-144. [doi: 10.1377/hlthaff.2008.0816] [Medline: 21209449]
2. Kwon S. Thirty years of national health insurance in South Korea: lessons for achieving universal health care coverage. Health Policy Plan 2009 Jan 12;24(1):63-71. [doi: 10.1093/heapol/czn037] [Medline: 19004861]
3. Song SO, Jung CH, Song YD, Park C, Kwon H, Cha BS, et al. Background and data configuration process of a nationwide population-based study using the korean national health insurance system. Diabetes Metab J 2014 Oct;38(5):395-403 [FREE Full text] [doi: 10.4093/dmj.2014.38.5.395] [Medline: 25349827]
4. Lim B. Korean medicine coverage in the National Health Insurance in Korea: present situation and critical issues. Integr Med Res 2013 Sep;2(3):81-88 [FREE Full text] [doi: 10.1016/j.imr.2013.06.004] [Medline: 28664058]
5. Korea Ministry of Employment and Labor. Korea Occupational Safety and Health Act of Korea. URL: http://www.moleg.go.kr/english [accessed 2018-08-01]
6. Bui DDA, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc 2014 Sep 01;21(5):850-857 [FREE Full text] [doi: 10.1136/amiajnl-2013-002411] [Medline: 24578357]
7. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc 2012 Sep 01;19(5):913-916 [FREE Full text] [doi: 10.1136/amiajnl-2011-000607] [Medline: 22291166]
8. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012 Sep 01;19(5):833-839 [FREE Full text] [doi: 10.1136/amiajnl-2012-000820] [Medline: 22542812]
9. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language processing: a methodical review. J Am Med Inform Assoc 2020 Mar 01;27(3):457-470 [FREE Full text] [doi: 10.1093/jamia/ocz200] [Medline: 31794016]
10. Sorin V, Barash Y, Konen E, Klang E. Deep Learning for Natural Language Processing in Radiology-Fundamentals and a Systematic Review. J Am Coll Radiol 2020 May;17(5):639-648. [doi: 10.1016/j.jacr.2019.12.026] [Medline: 32004480]
11. Leyh-Bannurah S, Tian Z, Karakiewicz PI, Wolffgang U, Sauter G, Fisch M, et al. Deep Learning for Natural Language Processing in Urology: State-of-the-Art Automated Extraction of Detailed Pathologic Prostate Cancer Data From Narratively Written Electronic Health Records. JCO Clinical Cancer Informatics 2018 Dec(2):1-9. [doi: 10.1200/cci.18.00080]
12. Xu J, Li Z, Wei Q, Wu Y, Xiang Y, Lee H, et al. Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text. BMC Med Inform Decis Mak 2019 Dec 05;19(Suppl 5):236 [FREE Full text] [doi: 10.1186/s12911-019-0937-2] [Medline: 31801529]
13. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak 2017 Jul 05;17(Suppl 2):67 [FREE Full text] [doi: 10.1186/s12911-017-0468-7] [Medline: 28699566]
14. Yang X, Lyu T, Li Q, Lee C, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak 2019 Dec 05;19(Suppl 5):232 [FREE Full text] [doi: 10.1186/s12911-019-0935-4] [Medline: 31801524]
15. Borjali A, Magnéli M, Shin D, Malchau H, Muratoglu OK, Varadarajan KM. Natural language processing with deep learning for medical adverse event detection from free-text medical narratives: A case study of detecting total hip replacement dislocation. Comput Biol Med 2021 Feb;129:104140. [doi: 10.1016/j.compbiomed.2020.104140] [Medline: 33278631]
16. Koleck T, Dreisbach C, Bourne P, Bakken S. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc 2019 Apr 01;26(4):364-379 [FREE Full text] [doi: 10.1093/jamia/ocy173] [Medline: 30726935]
17. LeCun Y, Bengio Y, Hinton G. Deep Learning. Cambridge, MA: MIT Press; 2016:-44.
18. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv 2013 Jan 16 [FREE Full text]
19. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv 2018 Oct 11.
20. Clark K, Luong MT, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. arXiv 2020 Mar 23.

21. Park J. KoELECTRA: Pretrained ELECTRA Model for Korean. GitHub repository. 2020. URL: https://github.com/monologg/KoELECTRA [accessed 2020-11-18]
22. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998 Nov;86(11):2278-2324. [doi: 10.1109/5.726791]
23. 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]
24. American Thyroid Association. Thyroid Function Tests. 2019. URL: https://www.thyroid.org/wp-content/uploads/patients/brochures/FunctionTests_brochure [accessed 2020-12-28]
25. National Cancer Institute. Thyroid Cancer Treatment (Adult). 2020. URL: https://www.cancer.gov/types/thyroid/patient/thyroid-treatment-pdq [accessed 2020-12-28]
26. Tsarfaty R, Seddah D, Kübler S, Nivre J. Parsing Morphologically Rich Languages: Introduction to the Special Issue. 2013 Mar. URL: https://aclanthology.org/J13-1003.pdf [accessed 2021-09-08]
27. Seddah D, Kübler S, Tsarfaty R. Introducing the spmrl 2014 shared task on parsing morphologically-rich languages. 2014. URL: https://aclanthology.org/W14-6111.pdf [accessed 2021-09-08]
28. Lee YW, Yoo YH. MeCab-Ko-Dic, A Piece of Silver Coin Project. URL: http://eunjeon.blogspot.com/ [accessed 2020-09-15]
29. Kudo T, Yamamoto K, Matsumoto Y. Applying Conditional Random Fields to Japanese Morphological Analysis. URL: https://aclanthology.org/W04-3230.pdf [accessed 2021-09-08]
30. Park E, Cho S. KoNLPy: Korean natural language processing in Python. In: Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology. 2014 Presented at: 26th Annual Conference on Human & Cognitive Language Technology; October 2014; Chuncheon, Korea p. 133-136.
31. Schuster M, Nakajima K. Japanese and Korean voice search. New York, NY: IEEE; 2012 Mar 25 Presented at: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 25-30, 2012; Kyoto, Japan p. 5149-5152. [doi: 10.1109/ICASSP.2012.6289079]
32. Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. arXiv 2015 Aug 31.
33. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv 2016 Sep 26.
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 2017 Presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017); December 4-9, 2017; Long Beach, CA p. 5998-6008 URL: https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
35. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on Machine Learning. 2015 Jun 01 Presented at: 32nd International Conference on Machine Learning; 2015; Lille, France p. 448-456 URL: http://proceedings.mlr.press/v37/ioffe15.pdf
36. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. New York, NY: IEEE; 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV p. 770-778. [doi: 10.1109/cvpr.2016.90]
37. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. 2016 Jun Presented at: Proceedings of NAACL-HLT 2016; June 12-17, 2016; San Diego, CA p. 1480-1489 URL: https://aclanthology.org/N16-1174.pdf [doi: 10.18653/v1/n16-1174]
38. Wikipedia:Database Download. URL: https://dumps.wikimedia.org/kowiki/ [accessed 2020-06-25]
39. Song X, Salcianu A, Song Y, Dopson D, Zhou D. Linear-Time WordPiece Tokenization. arXiv 2020 Dec 31.
40. 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 1;39(6):1137-1149. [doi: 10.1109/tpami.2016.2577031]
41. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. New York, NY: IEEE; 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV.
42. Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv 2019 May 24.
43. Young T, Hazarika D, Poria S, Cambria E. Recent Trends in Deep Learning Based Natural Language Processing [Review Article]. IEEE Comput. Intell. Mag 2018 Aug;13(3):55-75. [doi: 10.1109/mci.2018.2840738]
44. Nair V, Hinton G. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. Madison, WI: Omnipress; 2010 Jan 01 Presented at: 27th International Conference on Machine Learning; June 21-24, 2010; Haifa, Israel URL: https://www.cs.toronto.edu/~fritz/absps/reluICML.pdf
45. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 2014 Jan 01;15(1):1929-1958 [FREE Full text]
46. Van Laarhoven T. L2 regularization versus batch and weight normalization. arXiv 2017 Jun 16.
47. Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv 2014 Dec 22.
48. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv 2016 Mar 14 [FREE Full text]
49. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019 Presented at: 33rd Conference on Neural

Information Processing Systems (NeurIPS 2019); December 2, 2019; Vancouver, Canada p. 8026-8037 URL: https://research.fb.com/wp-content/uploads/2019/12/PyTorch-An-Imperative-Style-High-Performance-Deep-Learning-Library.pdf

50. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. Transformers: State-of-the-art natural language processing. Stroudsburg, PA: Association for Computational Linguistics; 2020 Oct Presented at: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; November 16-20, 2020; Virtual p. 38-45 URL: https://aclanthology.org/2020.emnlp-demos.6.pdf
51. Lin H, Rong R, Gao X, Revanna K, Zhao M, Bajic P, et al. Disease correlation network: a computational package for identifying temporal correlations between disease states from Large-Scale longitudinal medical records. JAMIA Open 2019 Oct;2(3):353-359 [FREE Full text] [doi: 10.1093/jamiaopen/ooz031] [Medline: 31984368]
52. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, et al. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun 2014 Jun 24;5(1):4022 [FREE Full text] [doi: 10.1038/ncomms5022] [Medline: 24959948]
53. Goh K, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci U S A 2007 May 22;104(21):8685-8690 [FREE Full text] [doi: 10.1073/pnas.0701361104] [Medline: 17502601]
54. Feezer LW. Theories concerning the causation of disease. Am J Public Health (N Y) 1921 Oct;11(10):908-912. [doi: 10.2105/ajph.11.10.908] [Medline: 18010574]
55. Najman JM. Theories of disease causation and the concept of a general susceptibility: A review. Social Science & Medicine. Part A: Medical Psychology & Medical Sociology 1980 Jan;14(3):231-237. [doi: 10.1016/s0271-7123(80)91733-2]
56. Thagard P. Explaining disease: Correlations, causes, and mechanisms. Minds and Machines 1998 Feb;8(1):61-78 [FREE Full text]
57. Broadbent A. Causation and models of disease in epidemiology. Stud Hist Philos Biol Biomed Sci 2009 Dec;40(4):302-311. [doi: 10.1016/j.shpsc.2009.09.006] [Medline: 19917489]

Abbreviations

AUC: area under the curve
BERT: bidirectional encoder representations from transformers
BN: batch normalization
CBOW: continuous bag of words
CNN: convolution neural network
DL: deep learning
ELECTRA: efficiently learning an encoder that classifies token replacements accurately
EMR: electronic medical record
LSTM: long short-term memory
NLP: natural language processing
OOV: out of vocabulary
RNN: recurrent neural network
ROC: receiver operating characteristic
SCENT: static and contextualized ensemble NLP network
SGNS: skip-gram with negative sampling

Edited by G Eysenbach; submitted 06.05.21; peer-reviewed by PP Zhao, V Gupta, V Dornauer; comments to author 31.05.21; revised version received 15.07.21; accepted 02.08.21; published 21.09.21.

Please cite as:
Shin D, Kam HJ, Jeon MS, Kim HY
Automatic Classification of Thyroid Findings Using Static and Contextualized Ensemble Natural Language Processing Systems: Development Study
JMIR Med Inform 2021;9(9):e30223
URL: https://medinform.jmir.org/2021/9/e30223
doi:10.2196/30223
PMID:34546183

Viewpoint

Defining Patient-Oriented Natural Language Processing: A New Paradigm for Research and Development to Facilitate Adoption and Use by Medical Experts

Abeed Sarker[1], PhD; Mohammed Ali Al-Garadi[1], PhD; Yuan-Chi Yang[1], PhD; Jinho Choi[2], PhD; Arshed A Quyyumi[3], MD; Greg S Martin[4], MSc, MD

[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA, United States

[2]Department of Computer Science, College of Arts and Sciences, Emory University, Atlanta, GA, United States

[3]Emory Clinical Cardiovascular Institute, Division of Cardiology, Department of Medicine, School of Medicine, Emory University, Atlanta, GA, United States

[4]Predictive Health Institute and Center for Health Discovery and Well Being, Department of Medicine, School of Medicine, Emory University, Atlanta, GA, United States

Corresponding Author:
Abeed Sarker, PhD
Department of Biomedical Informatics
School of Medicine
Emory University
101 Woodruff Circle
Office 4101
Atlanta, GA, 30322
United States
Phone: 1 404 712 0055
Email: abeed@dbmi.emory.edu

Abstract

The capabilities of natural language processing (NLP) methods have expanded significantly in recent years, and progress has been particularly driven by advances in data science and machine learning. However, NLP is still largely underused in patient-oriented clinical research and care (POCRC). A key reason behind this is that clinical NLP methods are typically developed, optimized, and evaluated with narrowly focused data sets and tasks (eg, those for the detection of specific symptoms in free texts). Such research and development (R&D) approaches may be described as *problem oriented*, and the developed systems perform specialized tasks well. As standalone systems, however, they generally do not comprehensively meet the needs of POCRC. Thus, there is often a gap between the capabilities of clinical NLP methods and the needs of patient-facing medical experts. We believe that to increase the practical use of biomedical NLP, future R&D efforts need to be broadened to a new research paradigm—one that explicitly incorporates characteristics that are crucial for POCRC. We present our viewpoint about 4 such interrelated characteristics that can increase NLP systems' suitability for POCRC (3 that represent NLP system properties and 1 associated with the R&D process)—(1) interpretability (the ability to explain system decisions), (2) patient centeredness (the capability to characterize diverse patients), (3) customizability (the flexibility for adapting to distinct settings, problems, and cohorts), and (4) multitask evaluation (the validation of system performance based on multiple tasks involving heterogeneous data sets). By using the NLP task of clinical concept detection as an example, we detail these characteristics and discuss how they may result in the increased uptake of NLP systems for POCRC.

(JMIR Med Inform 2021;9(9):e18471) doi:10.2196/18471

KEYWORDS

natural language processing; text mining; patient-centered care; evidence-based medicine; medical informatics

Introduction

Health informatics is an emerging interdisciplinary field that has undergone considerable evolution over recent years. This evolution has largely been driven by the availability of big data and progress in artificial intelligence, machine learning, and data science [1]. Big data from electronic health records (EHRs) have enabled researchers to train and execute neural network–based machine learning (eg, deep learning) algorithms for targeted problems, which have sometimes achieved performances that are comparable to those of human experts [2,3]. Clinical natural language processing (NLP)—one of the most complex subfields of health informatics—has also undergone rapid progress recently, which has been propelled by advanced machine learning, including deep learning [4] and text representation methods [5,6]. Clinical NLP holds particular promise for improving evidence-based, patient-oriented clinical research and care (POCRC), since significant volumes of knowledge regarding patients and research evidence are encapsulated in the form of free text [7,8]. Patient-centered medicine and patient-oriented research focus on the unique needs and characteristics of patients in addition to the specialized skills of domain experts and the best available research evidence [9-13]. Due to its emphasis on outcomes that are important to patients, the POCRC model has been suggested to be superior in terms of quality compared to disease-oriented models, which focus on surrogate end points such as laboratory measurements and physical signs [13-17]. There has therefore been a continuous push, particularly in the practice of evidence-based medicine, to promote POCRC.

NLP tools and methods are traditionally optimized and evaluated based on their abilities to perform specialized, problem-specific, site-specific technical tasks. Such methods typically lack the capabilities to go beyond the problems that they are developed for and are unable to describe the relevant diverse characteristics of individual patients or help medical experts with patient-oriented decision-making. For example, studies on the fundamental NLP task of clinical concept detection (ie, concepts from EHRs or other sources) are typically designed to detect or extract small sets of disease-specific or problem-specific homogeneous concepts and are evaluated intrinsically via metrics such as accuracy and the F-measure. Such concepts, for example, include health conditions such as obesity [18], bleeding [19], and drug reactions [20] and behavioral patterns such as tobacco [21] and alcohol [22] use. Velupillai et al [23] explained that although such systems may show high performances in intrinsic evaluation, they may have reduced value at the higher patient level. When the abovementioned problem-oriented NLP models are viewed through the lens of the well-defined model of patient-centered health care [9], they appear to be analogous to disease-oriented, evidence-based medicine models, as they focus on a particular disease or problem instead of holistically taking patients into account. Such problem-oriented NLP research and development (R&D) has resulted in the creation of state-of-the-art models for many clinical text processing tasks and is essential for incorporating NLP progress into health informatics. However, NLP methods' inability to meet the diverse requirements of medical experts has restricted their utility in POCRC. In a clinical scenario, particularly at the point of care, it is generally unrealistic to expect medical experts to customize and use multiple complex NLP methods to fully characterize patients based on the free-text information in patients' EHRs. As a consequence of these limitations, the transition of clinical NLP systems from their R&D environments to regular use by medical experts has been slow and limited [24,25]. By building on recent advances, clinical NLP R&D has the potential to progress from the use of disease- and problem-oriented models to the use of patient-oriented models, provided that the needs from an NLP perspective are clearly defined. The gap between the capabilities of NLP systems and the POCRC needs of medical experts may be due to the lack of specification regarding what a patient-oriented perspective for clinical NLP should comprise and how patient-oriented clinical NLP systems can complement traditional problem-oriented systems. There have been little to no formal schemes, definitions, or discussions in medical informatics literature about the aspects of patient-orientedness for NLP. Given the explosive recent advances in NLP, it is now crucial to establish the building blocks of the requirements of patient-oriented NLP, so that methodological research may be targeted to directly improve POCRC. In the following paragraphs, we attempt to formulate what aspects should be considered when developing patient-oriented NLP systems.

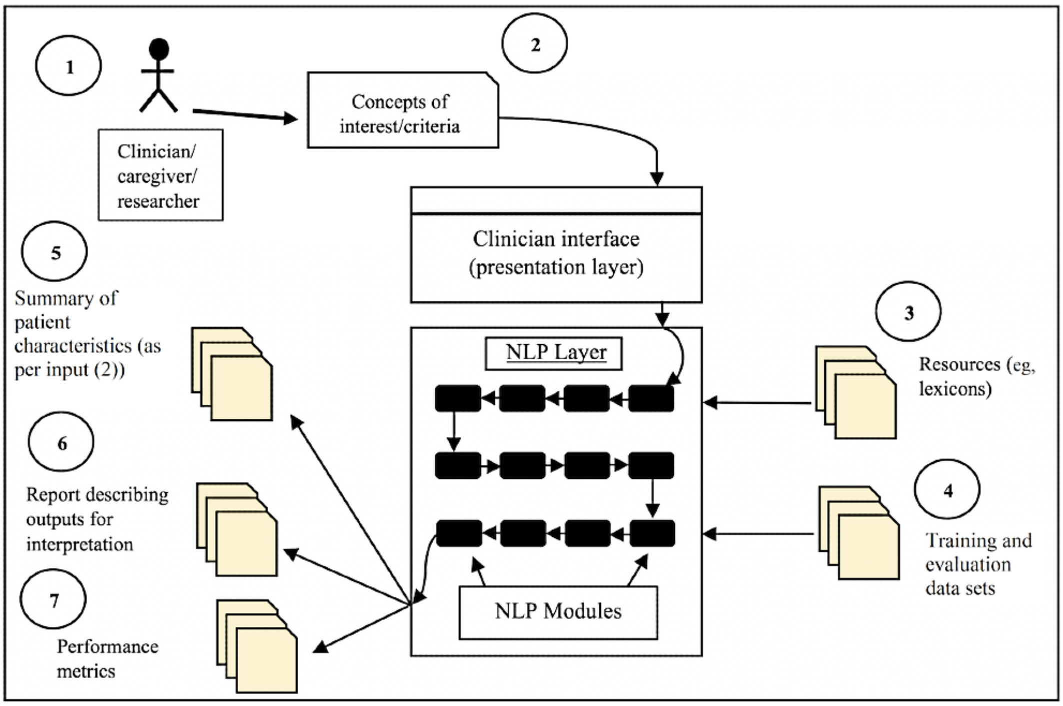
Key NLP Needs for POCRC

Interpretability as a Core System Component (Interpretability)

Recent advances in machine learning, particularly in deep learning, have resulted in their successful application to specific clinical tasks [26,27], and while most studies have relied on structured data from EHRs, some have used free-text information [4,28,29]. Some studies have even generated patient representations based on the nonlinear transformations of all encoded information in EHRs [30]. Despite the excellent results obtained by these systems in some cases, an obstacle to using these systems for POCRC—specifically when free text is involved—is the lack of interpretability. In fact, understanding how deep neural networks make their decisions is an area of active research in computer science [31,32]. Automation without interpretability means that the basis of a forecast or decision that is made by a system cannot be deciphered or explained by a medical expert. The inability to interpret the reasons behind automated systems' decisions results in the inability of patient-facing medical experts to communicate these reasons to patients for tasks such as shared decision-making.

When designing and developing clinical NLP systems, informaticians must consider interpretability as a necessary constraint. Black-box models may be effective for a given task, but unless the decisions of a system are traceable in the desired manner, their application may not evolve beyond the problem-specific task for which they were developed [33]. One method for potentially addressing this issue is integrating reporting mechanisms with machine learning models, so that the outputs of a task are not only predictions and numeric performance metrics but also modular reports that attempt to explain the reasons behind the predictions (eg, "which span of text in the note did the system think matched with concept X?" or "what were the top features that contributed to the system's decision?"). The hypothetical framework depicted in Figure 1 illustrates the generation of reports by a system alongside other outputs, such as performance metrics. Such reporting mechanisms are uncommon in current clinical NLP systems, as the focus of R&D is almost invariably on some type of problem-specific performance metric. This is one aspect in which involving clinical stakeholders in the development process is essential, as clinical interpretability needs may be distinct from mathematical or statistical interpretability needs [31,34].

Figure 1. An outline of a patient-oriented NLP framework illustrating (1) the ability of the caregiver to input the required criteria via an interface that is decoupled from the technical NLP modules and (2) outputs, including reports for ensuring interpretability. NLP: natural language processing.



Broadening the Scopes of Clinical NLP Systems (Patient Centeredness)

We envision that clinical NLP systems will see greater adoption and use by medical experts for POCRC if their scopes are broader and are centered on patients rather than problems. For example, in the task of clinical concept detection, the ideal NLP systems for domain experts (and, hence, the patients they serve) would be those designed to detect ad hoc clinical concepts in free text (as specified by the expert) rather than a set of homogenous concepts. Using the current problem-oriented NLP systems perhaps adds to the burden imposed on experts, such as the burden of the "4000 clicks per shift" [35] problem, and contributes to burnout [36]. In practice, patient-oriented researchers and caregivers require a holistic view of a patient, and from the perspective of clinical concept detection, such a representation of patients requires the detection of diverse information from patients' EHRs. Such information may range from typical concepts that past NLP research has focused on, such as diseases or symptoms, to atypical concepts such as descriptions of daily life interactions that affect the mental and physical well-being of a patient. This is perhaps the key reason why structured EHR data are preferred and are commonly used for patient-level analytical and predictive tasks. Such data present a varied set of information that, when combined, provides a detailed representation of a patient [37].

Future clinical NLP research that complements the existing advances in problem-based models should thus focus on developing frameworks that enable generalization at the patient level. For concept detection, this means enabling the specification of arbitrary clinical concepts of interest and detecting these concepts in the free-text portions of EHRs, which would result in the characterization of target patients based on these concepts. Since uncertainty is an inherent aspect of free text mining, instead of representing patient characteristics as binary variables, they can be represented by using continuous variables that represent the likelihood of a patient exhibiting specific characteristics (eg, the likelihood of viral exposure for a patient) [38]. Such a framework for concept detection can, for example, facilitate the construction of research cohorts or be used to identify eligible subjects for study enrollment based on the diverse subject information that is encoded in free text. We have seen some recent research in clinical NLP naturally evolve to take this approach to concept detection and patient characterization. For example, Stubbs et al [39] defined 13 variables, which involved diverse concepts that ranged from drug abuse to specified ranges of hemoglobin A_{1c} levels, for identifying patients who meet the selection criteria for a clinical trial. Although this approach to patient characterization via NLP was not explicitly described by the authors as *patient centered* and contrasted with typical problem-focused approaches, it represents a natural evolution toward patient-oriented NLP systems because its parameter flexibility can be used for practical tasks. Ideally, the technical complexities of the NLP algorithms for concept detection (or other purposes) should be decoupled from the interface that medical experts use, so that they may focus on specifying their patient-oriented needs (eg, ad hoc clinical concepts) without having to learn how to use multiple systems or how to execute such algorithms in multiple environments. Building NLP systems that are generalizable in such a manner is not trivial by any means, but we believe that the time is now right for designing and developing clinical NLP frameworks that incorporate such broader scopes.

Flexible Systems Are More Likely to Stand the Test of Time (Customizability)

A problem that has been plaguing clinical NLP systems is the lack of customizability and adaptability. Many systems are so specialized to the problem-specific task for which they were designed that substantial effort is needed to adapt them to other tasks or data sources [24,40]. The complexities of most clinical NLP systems, particularly those of recent systems that involve resource-heavy language models and intricate machine learning codes (eg, systems written in TensorFlow [41]), are difficult for medical experts with non-NLP educational backgrounds to comprehend. As such, even for very similar tasks, such experts cannot customize previously developed systems to address the needs of new studies. We suspect that in most cases clinician researchers and caregivers do not even consider the possibility of diving deep into system source codes (eg, those of potentially customizable tools such as the Clinical Language Annotation, Modeling, and Processing Toolkit [42]) and customizing them according to the specific needs of a study, as they are already burdened with information overload [43].

Clinical NLP systems should thus focus on simplicity and customizability. Incorporating these aspects into the R&D of clinical NLP systems is also not trivial. However, they may be achieved by adhering to typical software development best practices. This may include using layered architectures, in which complexities are hidden under simple interfaces that expose users to customizable options. Such an architecture is shown in Figure 1. In terms of clinical concept detection, the customizability of clinical NLP systems should enable medical experts to not only specify ad hoc concepts but also tune the system for different patient-oriented tasks (eg, cohort selection) by modifying system inputs, configurations, or parameters. Improving the customizability and simplicity of clinical NLP systems will undoubtedly increase their use in POCRC.

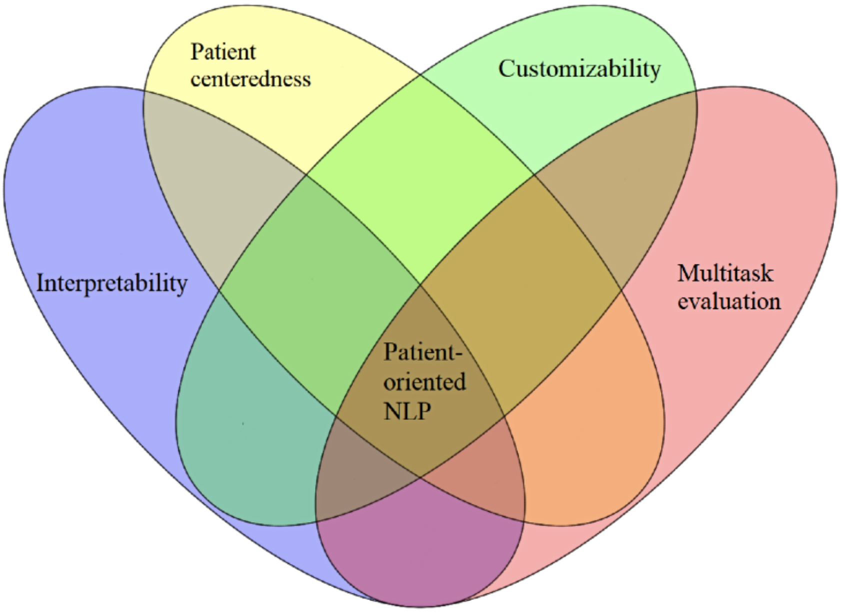
System Evaluations Using Multiple Data Sets With Heterogeneous Information (Multitask Evaluation)

System performance metrics obtained via evaluations based on a single data set can be misleading. Typical EHR-based free-text data sets are often constrained to small sets of patients with similar conditions, clinical settings, and social determinants, thereby causing systems that are built and evaluated based on such data sets to be overfit to the problem being studied [44]. Furthermore, the unique characteristics of the site from which the EHRs originated, such as the focus of the entity (eg, an urban children's hospital referral center) and the educational and training backgrounds of the note writers (eg, primary care physicians vs subspecialists), also influence how free text components are written. To gauge the true performances of clinical NLP methods, including performances associated with the three previously mentioned aspects, evaluations must be conducted based on multiple data sets with differing characteristic. The reuse utility of a system is substantially diminished if it is overfit to the characteristics of a specific data set. Reporting a system's performance metrics (eg, the F-measure for concept detection) based solely on intrinsic evaluations of such specialized data sets may also be potentially perilous, since future users may incorrectly assume that the system will exhibit similar performances on other data sets. Consequently, the evaluation of systems based on multiple data sets with distinct characteristics is imperative for ensuring the robustness of systems.

Conclusion

To facilitate the greater adoption of NLP in POCRC, R&D models need to build on problem-oriented approaches and transition to patient-oriented ones. In this paper, we outlined the fundamental characteristics of patient-oriented NLP system design and development. We discussed 4 interrelated factors (Figure 2) that are essential—(1) interpretability, (2) patient centeredness, (3) customizability, and (4) multitask evaluation. We believe that given the rapid recent advances in data science, it is time to initiate a new paradigm for NLP R&D—one with a patient-oriented focus—in order to increase the impact that NLP R&D has on health care. Such a paradigm shift will require overcoming many barriers, which include, but are not limited to, challenges posed by informal texts, diversities in health-related languages [24], the scarcity of annotated or labeled data, and difficulties that inhibit NLP systems' progress from processing texts to understanding them [45]. Recent advances in NLP, such as low-shot learning [46], have the potential to aid researchers with the development of systems that are patient-oriented and, consequently, increase the impact of NLP in health care. This paradigm shift will be necessarily incremental, as researchers will build on and improve initial systems over time.

Figure 2. The four foundational components of patient-oriented NLP. NLP: natural language processing.



Authors' Contributions

AS outlined the initial vision for this viewpoint paper. All coauthors contributed to the specification of the four factors discussed and contributed to the writing of this manuscript.

Conflicts of Interest

None declared.

References

1. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
2. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. Radiology 2018 Apr;287(1):313-322. [doi: 10.1148/radiol.2017170236] [Medline: 29095675]
3. 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]
4. 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]
5. Mikolov T, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. 2013 Dec Presented at: The 26th International Conference on Neural Information Processing Systems; December 5-10, 2013; Lake Tahoe, Nevada, USA p. 3111-3119.
6. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 2-7, 2019; Minneapolis, Minnesota p. 4171-4186. [doi: 10.18653/v1/N19-1423]
7. Simmons M, Singhal A, Lu Z. Text mining for precision medicine: Bringing structure to EHRs and biomedical literature to understand genes and health. Adv Exp Med Biol 2016;939:139-166 [FREE Full text] [doi: 10.1007/978-981-10-1503-8_7] [Medline: 27807747]
8. Alsawas M, Alahdab F, Asi N, Li DC, Wang Z, Murad MH. Natural language processing: use in EBM and a guide for appraisal. Evid Based Med 2016 Aug;21(4):136-138. [doi: 10.1136/ebmed-2016-110437] [Medline: 27284128]
9. The advanced medical home: A patient-centered physician-guided model of healthcare. American College of Physicians. 2005. URL: https://www.acponline.org/acp_policy/policies/adv_medicalhome_patient_centered_model_healthcare_2006.pdf [accessed 2021-08-24]
10. Evidence-Based Medicine Working Group. Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA 1992 Nov 04;268(17):2420-2425. [doi: 10.1001/jama.1992.03490170092032] [Medline: 1404801]

11. Greenhalgh T, Howick J, Maskrey N, Evidence Based Medicine Renaissance Group. Evidence based medicine: a movement in crisis? BMJ 2014 Jun 13;348:g3725 [FREE Full text] [doi: 10.1136/bmj.g3725] [Medline: 24927763]
12. Vandermause R, Barg FK, Esmail L, Edmundson L, Girard S, Perfetti AR. Qualitative methods in patient-centered outcomes research. Qual Health Res 2017 Feb;27(3):434-442. [doi: 10.1177/1049732316668298] [Medline: 27634294]
13. Sacristán JA. Patient-centered medicine and patient-oriented research: improving health outcomes for individual patients. BMC Med Inform Decis Mak 2013 Jan 08;13:6 [FREE Full text] [doi: 10.1186/1472-6947-13-6] [Medline: 23294526]
14. Godlee F. Outcomes that matter to patients. BMJ 2012 Jan 11;344(7839):e318. [doi: 10.1136/bmj.e318]
15. Green AR, Carrillo JE, Betancourt JR. Why the disease-based model of medicine fails our patients. West J Med 2002 Mar;176(2):141-143 [FREE Full text] [Medline: 11897746]
16. Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? Ann Intern Med 1996 Oct 01;125(7):605-613. [doi: 10.7326/0003-4819-125-7-199610010-00011] [Medline: 8815760]
17. de Grooth H, Parienti J, Oudemans-van Straaten HM. Should we rely on trials with disease- rather than patient-oriented endpoints? Intensive Care Med 2018 Apr;44(4):464-466 [FREE Full text] [doi: 10.1007/s00134-017-4859-0] [Medline: 28608246]
18. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. Appl Clin Inform 2016 Jul 20;7(3):693-706 [FREE Full text] [doi: 10.4338/ACI-2016-01-RA-0015] [Medline: 27452794]
19. Li R, Hu B, Liu F, Liu W, Cunningham F, McManus DD, et al. Detection of bleeding events in electronic health record notes using convolutional neural network models enhanced with recurrent neural network autoencoders: Deep learning approach. JMIR Med Inform 2019 Feb 08;7(1):e10788 [FREE Full text] [doi: 10.2196/10788] [Medline: 30735140]
20. Li F, Yu H. An investigation of single-domain and multidomain medication and adverse drug event relation extraction from electronic health record notes using advanced deep learning models. J Am Med Inform Assoc 2019 Jul 01;26(7):646-654 [FREE Full text] [doi: 10.1093/jamia/ocz018] [Medline: 30938761]
21. Hegde H, Shimpi N, Glurich I, Acharya A. Tobacco use status from clinical notes using natural language processing and rule based algorithm. Technol Health Care 2018;26(3):445-456. [doi: 10.3233/THC-171127] [Medline: 29614708]
22. Afshar M, Phillips A, Karnik N, Mueller J, To D, Gonzalez R, et al. Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. J Am Med Inform Assoc 2019 Mar 01;26(3):254-261. [doi: 10.1093/jamia/ocy166] [Medline: 30602031]
23. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, et al. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform 2018 Dec;88:11-19 [FREE Full text] [doi: 10.1016/j.jbi.2018.10.005] [Medline: 30368002]
24. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc 2017 Sep 01;24(5):986-991 [FREE Full text] [doi: 10.1093/jamia/ocx039] [Medline: 28419261]
25. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: Leveraging the unstructured content in electronic health records. Front Med (Lausanne) 2019 Apr 17;6:66 [FREE Full text] [doi: 10.3389/fmed.2019.00066] [Medline: 31058150]
26. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 2018 May 08;1:18 [FREE Full text] [doi: 10.1038/s41746-018-0029-1] [Medline: 31304302]
27. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inform Decis Mak 2018 Dec 12;18(Suppl 4):122 [FREE Full text] [doi: 10.1186/s12911-018-0677-8] [Medline: 30537977]
28. 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]
29. Zhu R, Tu X, Huang J. Using deep learning based natural language processing techniques for clinical decision-making with EHRs. In: Dash S, Acharya B, Mittal M, Abraham A, Kelemen A, editors. Deep Learning Techniques for Biomedical and Health Informatics. Cham, Switzerland: Springer International Publishing; 2020:257-295.
30. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 May 17;6:26094 [FREE Full text] [doi: 10.1038/srep26094] [Medline: 27185194]
31. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018 Feb;73:1-15 [FREE Full text] [doi: 10.1016/j.dsp.2017.10.011]
32. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. 2019 Jul 17 Presented at: The 33rd AAAI Conference on Artificial Intelligence; January 27 to February 1, 2019; Honolulu, Hawaii, USA p. 3681-3688. [doi: 10.1609/aaai.v33i01.33013681]
33. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018 Apr;15(141):20170387 [FREE Full text] [doi: 10.1098/rsif.2017.0387] [Medline: 29618526]
34. Hohman FM, Kahng M, Pienta R, Chau DH. Visual analytics in deep learning: An interrogative survey for the next frontiers. IEEE Trans Vis Comput Graph 2018 Jun 04. [doi: 10.1109/TVCG.2018.2843369] [Medline: 29993551]

35. Hill RG, Sears LM, Melanson SW. 4000 clicks: a productivity analysis of electronic medical records in a community hospital ED. Am J Emerg Med 2013 Nov;31(11):1591-1594. [doi: 10.1016/j.ajem.2013.06.028] [Medline: 24060331]
36. Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. JAMA 2018 Jan 02;319(1):19-20. [doi: 10.1001/jama.2017.19198] [Medline: 29261830]
37. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208 [FREE Full text] [doi: 10.1093/jamia/ocw042] [Medline: 27189013]
38. Sarker A, Klein AZ, Mee J, Harik P, Gonzalez-Hernandez G. An interpretable natural language processing system for written medical examination assessment. J Biomed Inform 2019 Oct;98:103268. [doi: 10.1016/j.jbi.2019.103268] [Medline: 31421211]
39. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc 2019 Nov 01;26(11):1163-1171 [FREE Full text] [doi: 10.1093/jamia/ocz163] [Medline: 31562516]
40. Johnson SB, Adekkanattu P, Campion TR, Flory J, Pathak J, Patterson OV, et al. From sour grapes to low-hanging fruit: A case study demonstrating a practical strategy for natural language processing portability. AMIA Jt Summits Transl Sci Proc 2018 May 18;2017:104-112 [FREE Full text] [Medline: 29888051]
41. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: A system for large-scale machine learning. 2016 Presented at: The 12th USENIX Conference on Operating Systems Design and Implementation; November 2-4, 2016; Savannah, Georgia, USA p. 265-283 URL: https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf [doi: 10.5555/3026877.3026899]
42. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018 Mar 01;25(3):331-336 [FREE Full text] [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
43. Klerings I, Weinhandl AS, Thaler KJ. Information overload in healthcare: too much of a good thing? Z Evid Fortbild Qual Gesundhwes 2015;109(4-5):285-290. [doi: 10.1016/j.zefq.2015.06.005] [Medline: 26354128]
44. Oleynik M, Kugic A, Kasáč Z, Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J Am Med Inform Assoc 2019 Nov 01;26(11):1247-1254 [FREE Full text] [doi: 10.1093/jamia/ocz149] [Medline: 31512729]
45. Sheikhalishahi S, Miotto R, Dudley JT, Lavelli A, Rinaldi F, Osmani V. Natural language processing of clinical notes on chronic diseases: Systematic review. JMIR Med Inform 2019 Apr 27;7(2):e12239 [FREE Full text] [doi: 10.2196/12239] [Medline: 31066697]
46. Xia C, Zhang C, Zhang J, Liang T, Peng H, Yu PS. Low-shot learning in natural language processing. 2020 Presented at: 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI); October 28-31, 2020; Atlanta, Georgia, USA p. 185-189. [doi: 10.1109/cogmi50398.2020.00031]

Abbreviations

EHR: electronic health record
NLP: natural language processing
POCRC: patient-oriented clinical research and care
R&D: research and development

Edited by R Kukafka, G Eysenbach; submitted 28.02.20; peer-reviewed by L Li, N Karnik, MG Kim, V Foufi; comments to author 17.09.20; revised version received 13.02.21; accepted 02.08.21; published 28.09.21.

Please cite as:
Sarker A, Al-Garadi MA, Yang YC, Choi J, Quyyumi AA, Martin GS
Defining Patient-Oriented Natural Language Processing: A New Paradigm for Research and Development to Facilitate Adoption and Use by Medical Experts
JMIR Med Inform 2021;9(9):e18471
URL: https://medinform.jmir.org/2021/9/e18471
doi:10.2196/18471
PMID:34581670

Original Paper

Prediction of Critical Care Outcome for Adult Patients Presenting to Emergency Department Using Initial Triage Information: An XGBoost Algorithm Analysis

Hyoungju Yun[1], MISM; Jinwook Choi[1,2,3], MD, PhD; Jeong Ho Park[4,5], MD

[1]Interdisciplinary Program of Medical Informatics, College of Medicine, Seoul National University, Seoul, Republic of Korea
[2]Department of Biomedical Engineering, College of Medicine, Seoul National University, Seoul, Republic of Korea
[3]Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, Seoul, Republic of Korea
[4]Department of Emergency Medicine, College of Medicine, Seoul National University, Seoul, Republic of Korea
[5]Laboratory of Emergency Medical Services, Biomedical Research Institute, Seoul National University Hospital, Seoul, Republic of Korea

Corresponding Author:
Jinwook Choi, MD, PhD
Institute of Medical and Biological Engineering
Medical Research Center
Seoul National University
103 Daehak-Ro, Jongno-Gu
Seoul, 03080
Republic of Korea
Phone: 82 2 2072 3421
Email: jinchoi@snu.ac.kr

Abstract

Background: The emergency department (ED) triage system to classify and prioritize patients from high risk to less urgent continues to be a challenge.

Objective: This study, comprising 80,433 patients, aims to develop a machine learning algorithm prediction model of critical care outcomes for adult patients using information collected during ED triage and compare the performance with that of the baseline model using the Korean Triage and Acuity Scale (KTAS).

Methods: To predict the need for critical care, we used 13 predictors from triage information: age, gender, mode of ED arrival, the time interval between onset and ED arrival, reason of ED visit, chief complaints, systolic blood pressure, diastolic blood pressure, pulse rate, respiratory rate, body temperature, oxygen saturation, and level of consciousness. The baseline model with KTAS was developed using logistic regression, and the machine learning model with 13 variables was generated using extreme gradient boosting (XGB) and deep neural network (DNN) algorithms. The discrimination was measured by the area under the receiver operating characteristic (AUROC) curve. The ability of calibration with Hosmer–Lemeshow test and reclassification with net reclassification index were evaluated. The calibration plot and partial dependence plot were used in the analysis.

Results: The AUROC of the model with the full set of variables (0.833-0.861) was better than that of the baseline model (0.796). The XGB model of AUROC 0.861 (95% CI 0.848-0.874) showed a higher discriminative performance than the DNN model of 0.833 (95% CI 0.819-0.848). The XGB and DNN models proved better reclassification than the baseline model with a positive net reclassification index. The XGB models were well-calibrated (Hosmer-Lemeshow test; *P*>.05); however, the DNN showed poor calibration power (Hosmer-Lemeshow test; *P*<.001). We further interpreted the nonlinear association between variables and critical care prediction.

Conclusions: Our study demonstrated that the performance of the XGB model using initial information at ED triage for predicting patients in need of critical care outperformed the conventional model with KTAS.

(JMIR Med Inform 2021;9(9):e30770) doi:10.2196/30770

KEYWORDS

triage; critical care; prediction; XGBoost; explainable machine learning; interpretable artificial intelligence; machine learning; algorithm; prediction; outcome; emergency; triage; classify; prioritize; risk; model

Introduction

Overcrowding in the emergency department (ED) has become a major worldwide health care problem [1-3]. Therefore, most EDs have a triage to manage growing patient volumes [2,4,5]. ED triage is the first risk assessment for prioritizing patients at high risk and determining the course of ED care for patients [5-8]. It is vital to accurately identify patients who need immediate care at triage and provide rapid care to patients in ED since delay in care may result in increased morbidity and mortality for many clinical conditions [2,4,5,7,9,10].

Five-level triage systems, including the Canadian Triage and Acuity Scale (CTAS), Manchester Triage System (MTS), and emergency severity index (ESI), are widely used [2,8,9]. The Korean Triage and Acuity Scale (KTAS) was developed in 2012 based on CTAS and has been used nationally as the ED triage tool in Korea since 2016 [11-13]. Although five-level triage systems are well established in ED, they need to be improved because they heavily rely on healthcare providers' subjective judgment, resulting in high variability [5,7-10,12].

Machine learning algorithms such as extreme gradient boosting (XGB) and deep neural networks (DNNs) have the advantage of fitting nonlinear relationships between predictors and outcomes in large data sets [10,14-17]. Recent literature has shown machine learning prediction models using triage information perform better than the baseline model using the conventional approach of the five-level triage score for screening ED patients at risk of hospitalization, intensive care unit (ICU) admission, mortality, and critical care, which is defined as the combined outcome of ICU admission and mortality [3,6-10,12,17-20].

Clinical prediction models should be characterized by discrimination, which indicates how well the model differentiates patients who will have an event from those who will not, and by calibration, which refers to the agreement between predictions and the observed outcome [20-23]. Systematic reviews have reported that machine learning model studies for clinical predictions almost always assessed discriminative performance using the area under the receiver operating characteristic (AUROC) curve, and the reliability of risk prediction, namely calibration, was rarely evaluated [24-27]. In most of the previous studies for triage in ED, performance metrics pertaining to discriminating power were provided, but calibration, which assesses how close the prediction is to the true risk, was rarely reported. Raita et al provided the AUROC of ED triage prediction of critical care outcomes using four machine learning algorithms [9]. Kwon et al evaluated the discrimination of deep learning–based triage and acuity score model for critically ill patients [12]. Goto et al [10] investigated the discriminative performance of machine learning approaches for predicting critical care outcomes for patients with asthma and chronic obstructive pulmonary disease exacerbations in the ED. However, the calibration of the models for critical care outcomes was not included as a performance measure in the studies reviewed. Poorly calibrated prediction algorithm models can be misleading, which may result in incorrect and potentially harmful clinical decisions [24,26-28]. Therefore, a study including a performance evaluation of calibration in the prediction model for patients with a critical illness at triage in ED is required.

Moreover, no study has investigated the interpretability of machine learning models for the triage in ED to date. The interpretability of machine learning is defined as the degree to which the machine learning user understands and interprets the prediction made by a machine learning model [14-16]. The lack of interpretation is the barrier to establishing clinicians' trust and the broader adoption of machine learning models in clinical practices [14,15,29]. Explaining the justification of prediction outcomes of the machine learning algorithm model ensuring that the model makes the right predictions for the right reasons is required to enhance clinicians' buy-in [14-16,29]. Therefore, in this study, we apply the partial dependence plot (PDP), a global model-agnostic technique for explaining the relationship between predictors and prediction results, to investigate the interpretability of machine learning prediction for clinical care in ED [15,16].

We developed and validated the machine learning prediction model for critical care outcomes using routinely available triage information. We hypothesized that applying a machine learning algorithm to ED triage information could improve the performance of critical care outcome prediction for patients who visited an ED compared with the baseline KTAS model using logistic regression.

Methods

Study Design, Setting, and Data Source

This was a retrospective study of patients that visited the emergency department of an urban tertiary-care academic center with an annual census of about 70,000 from January 1, 2016, and December 31, 2018. We collected the demographics (age and gender), mode of ED arrival, the time interval between onset and ED visit, reason of ED visit, chief complaint, initial vital sign measurements, KTAS score, and disposition results (ED results and admission results). All data were acquired from the Korean National Emergency Department Information System.

Study Population

We considered adult patients (aged ≥18 years) who visited an ED during the study period. We excluded patients who did not need clinical outcomes prediction at triage, that is, cardiac arrest or death upon ED arrival. Furthermore, we excluded patients transferred to another hospital or those with uncompleted care because it was impossible to ascertain their ED results. Patients with missing or invalid information at triage were not included (Table S1, Multimedia Appendix 1).

Outcome

The primary outcome in this study was critical care outcome, defined as the composite of direct admission to ICU or in-hospital mortality following previous studies [4,7,9].

Variables and Preprocessing

For the prediction of critical care, we included a total of 13 variables: age, gender, mode of ED arrival, the time interval between onset and ED arrival, reason of ED visit, chief complaint, systolic blood pressure (SBP), diastolic blood pressure (DBP), pulse rate (PR), respiratory rate (RR), body temperature (BT), oxygen saturation, and level of consciousness namely, alert, verbal, painful, and unresponsive (AVPU). The mode of ED arrival was categorized into two options as either ambulance use or not. The reason for the ED visit had two values, either illness or injury. The chief complaints, which were based on the Unified Medical Language System (UMLS), were selected from the list of 547 codes. The preprocessing details for the variables are described in Multimedia Appendix 1 (Table S1).

Model Development

The prediction model of critical outcome was developed by using two modern prediction algorithms: XGB and DNN.

XGB algorithm is a cutting-edge machine learning application of gradient boosting mechanisms [3,8,9,30]. The gradient boosting is an ensemble algorithm with which new trees focus on adjusting errors produced by the previous tree models [8,30-32]. We implemented the XGB model on the training set using five-fold cross-validation. The maximum depth of five and a learning rate of 0.1 were selected from grid search for tuning hyperparameter (Table S2, Multimedia Appendix 1). For a DNN algorithm that equips the learning mechanism to fit nonlinear relationships and high order interactions, [5,10,20,33], we used three hidden layers selected from the grid search: (1) a rectified linear unit as the activation function; (2) an adaptive moment estimation as the optimizer; (3) a drop-out rate of 10%, zero value for lambda, and binary cross-entropy as the loss function (Table S2, Multimedia Appendix 1).

Random sampling was applied to split the entire data set into training (80%) and validation sets (20%). The performance of the prediction model was evaluated in the validation data set.

Statistical Analysis

For the characteristics of the study population according to critical care, a two-tailed *t* test or Mann–Whitney *U* test was conducted for the continuous variables, and the chi-square test or Fisher's exact test was performed for the categorical variables.

The discriminating power as a primary measure was evaluated by AUROC, which refers to how well the model differentiates those at a higher risk of having an event from those at lower risk [17,21]. We used the DeLong test to compare AUROC between models [9]. Reclassification improvement was evaluated using the net reclassification index (NRI) [9,10,21]. The NRI quantifies how well a new model reclassifies subjects compared with the reference model [9,10,21]. Model calibration was assessed with the Hosmer-Lemeshow test, a goodness-of-fit measure for prediction models of binary outcomes [20,21,23,34]. Furthermore, the calibration was depicted on a reliability diagram to represent the relationship between predicted probability and observed outcomes [17,20,21,23,34]. The perfect calibration should be in the 45-degree line [17,23,34]. The sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) were reported on performance metrics. We used a sensitivity cutoff point of 85% for the illustration of performance.

The variable importance of each prediction model was assessed and determined using the approach of permutation variable importance, which computes the importance by measuring the decrease of model prediction performance (AUROC) when each variable is permuted [35-38].

Finally, for the best prediction model, the PDP was visualized for both the direction and effect size of each variable after averaging out the effect of the other predictors in the model [38-40]. More concretely, the partial dependence by calculating the marginal effect of a single variable on the prediction outcome demonstrates whether the association between a variable and the prediction response is linear or nonlinear [15,40,41].

A two-tailed *P* value of <.05 was considered statistically significant, and a 95% CI was provided. All analyses were performed using the R software (version 3.6.1, R Foundation for Statistical Computing).

Ethics Statement

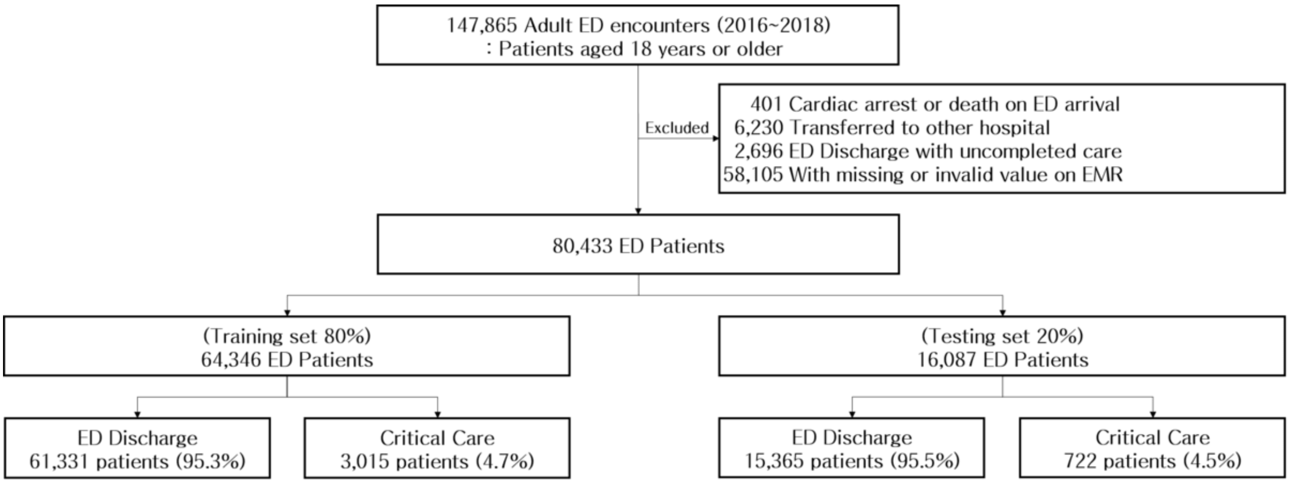
The Institutional Review Board of Seoul National University Hospital approved this study, and they waived the requirement for consent. All methods were performed in accordance with the relevant guidelines and regulations.

Results

Characteristics of Study Subjects

There were 147,865 adult ED encounters from January 1, 2016, to December 31, 2018. After excluding patients with cardiac arrest or death upon ED arrival (n=401), those transferred to another hospital (n=6230), discharged with uncompleted care (n=2696), and with missing or invalid values (n=58,105), a total of 80,433 ED adult patients were included in this study, with 3737 (4.6%) of them identified as experiencing critical care (Figure 1).

Figure 1. Study population. ED: emergency department. EMR: electronic medical record.



The study population of this study was split into two samples: (1) a training data set, comprising 80% of the data set, with 64,346 patients and containing 3015 (4.7%) critical care patients, and (2) a validation data set, consisting of the remaining 20% of the data set, with 16,807 patients, including 722 (4.5%) of them ascertained as receiving critical care. The characteristics of the training and validation data sets were not significantly different (Table S3, Multimedia Appendix 1).

The characteristics of the ED patients according to the study outcome are presented in Table 1. Critically ill patients were more likely to be female, older, call EMS, and have a higher proportion of illness than those without critical care. The time interval between onset and ED arrival was not significantly different between patients with and without critical care. Initial vital signs and levels of consciousness were significantly different between the two groups. The most common chief complaint among critically ill patients was dyspnea and fever among those without critical care. The median of KTAS at ED triage was 2 points (emergent level) for the critical care group and 3 points (urgent level) for the noncritical care group. The ED length of stay of patients was 6.4 hrs in the critical care group and 4.0 hrs in the noncritical care group (Table 1).

Table 1. Baseline characteristics of adult emergency patients according to critical care.

Characteristic	Total (N=80,433)	ED[a] discharge (n=76,696)	Critical care (n=3737)	*P* value
Gender, n (%)				<.001
Male	39,210 (48.7)	37,010 (48.3)	2200 (58.9)	
Female	41,223 (51.3)	39,686 (51.7)	1537 (41.1)	
Age, median (IQR)	61.0 (46.0-73.0)	61.0 (45.0-72.0)	69.0 (58.0-77.0)	<.001
Interval between onset and ED arrival (hour), median (IQR)	23.9 (3.8-96.0)	23.9 (3.8-96.0)	23.1 (4.4-95.8)	.17
Mode of ED arrival (EMS[b] use), n (%)	19,264 (24.0)	17,162 (22.4)	2102 (56.2)	<.001
Reason for ED visit, n (%)				<.001
Illness	73,645 (91.6)	70,021 (91.3)	3624 (97.0)	
Injury	6788 (8.4)	6675 (8.7)	113 (3.0)	
Initial vital sign data, median (IQR)				
SBP[c], mmHg	141.0 (126.0-165.0)	142.0 (126.0-165.0)	133.0 (113.0-160.0)	<.001
DBP[d], mmHg	81.0 (72.0-92.0)	82.0 (72.0-92.0)	75.0 (63.0-88.0)	<.001
PR[e], beats/min	86.0 (74.0-101.0)	86.0 (74.0-101.0)	94.0 (77.0-112.0)	<.001
RR[f], breaths/min	18.0 (16.0-20.0)	18.0 (16.0-20.0)	20.0 (18.0-24.0)	<.001
BT[g], °C	36.5 (36.3-36.7)	36.5 (36.3-36.7)	36.5 (36.3-37.0)	<.001
SpO_2[h], %	97.0 (96.0-98.0)	97.0 (96.0-98.0)	97.0 (94.0-98.0)	<.001
Nonalert, n (%)	3592 (4.5)	2858 (3.7)	734 (19.6)	<.001
Chief complaint, n (%)				<.001
Dyspnea	7705 (9.6)	6793 (8.9)	912 (24.4)	
Fever	7275 (9.0)	6991 (9.1)	284 (7.6)	
Abdominal pain	5302 (6.6)	5136 (6.7)	166 (4.4)	
Chest pain	5042 (6.3)	4487 (5.9)	555 (14.9)	
Dizziness	3550 (4.4)	3505 (4.6)	45 (1.2)	
Others	51,559 (64.1)	49,784 (64.9)	1775 (47.5)	
KTAS[i] level, n (%)				<.001
1: Resuscitation	870 (1.1)	404 (0.5)	466 (12.5)	
2: Emergent	12,646 (15.7)	10,692 (13.9)	1954 (52.3)	
3: Urgent	47,977 (59.6)	46,702 (60.9)	1275 (34.1)	
4: Less urgent	16,637 (20.7)	16,599 (21.6)	38 (1.0)	
5: Nonurgent	2303 (2.9)	2299 (3.0)	4 (0.1)	
ED LOS[j] (hour), median (IQR)	4.1 (2.4-7.3)	4.0 (2.4-7.2)	6.4 (3.7-10.4)	<.001
ED disposition, n (%)				<.001
ED discharge	57,014 (70.9)	57,014 (74.3)	0 (0.0)	
Ward admission	19,123 (23.8)	18,630 (24.3)	493 (13.2)	
ICU[k] admission	3170 (3.9)	0 (0.0)	3170 (84.8)	
OR[l] admission	1080 (1.3)	1052 (1.4)	28 (0.7)	
ED mortality	46 (0.1)	0 (0.0)	46 (1.2)	
In-hospital mortality, n (%)	804 (1.0)	0 (0.0)	804 (21.5)	<.001

[a] ED: emergency department.

[b] EMS: emergency medical service.

[c] SBP: systolic blood pressure.

[d] DBP: diastolic blood pressure.

[e] PR: pulse rate.

[f] RR: respiratory rate.

[g] BT: body temperature.

[h] SpO_2: oxygen saturation.

[i] KTAS: Korean Triage and Acute Scale.

[j] LOS: length of stay.

[k] ICU: intensive care unit.

[l] OR: operating room.

Main Analysis

Classification results for the validation data set are presented in Table 2. While the baseline model with a single variable of KTAS had the lowest discriminative ability of AUROC 0.796 (95% CI 0.781-0.811), the machine learning models had higher discriminative ability. When using triage information, age, gender, mode of ED arrival, the time interval between onset and ED arrival, reason of ED visit, chief complaints, the six vital sign measurements, and level of consciousness, the XGB algorithm yielded a higher AUROC of 0.861 (95% CI 0.848-0.874) than DNN of 0.833 (95% CI 0.819-0.848) for the validation data set. The machine learning models achieved higher reclassification improvement over the reference model with positive NRI (*P*<.05). As Figure 2 depicted, the AUROCs between the models with the full set of variables and the baseline model were significantly different. (DeLong's test for the validation data set: *P*<.05) The XGB model showed good calibration (Hosmer–Lemeshow test for the validation data set: *P*>.05), and calibration of the DNN model was poor with *P*<.001. The calibration plots on the validation data set were illustrated in Figure 3. We selected the XGB model as the final model in this study, considering discrimination, net reclassification, and calibration.

The predictive performance metrics of the validation cohort, including sensitivity, specificity, PPV, and NPV, are presented in Table 3. The XGB and DNN model showed a higher sensitivity of 0.85 than the baseline model (0.65, 95% CI 0.61-0.68) with a cutoff at the level of KTAS 2. As a trade-off, the specificity of the conventional model using a single variable of KTAS had a higher specificity of 0.85 (95% CI 0.84-0.86) than that of the XGB model at 0.71 (95% CI 0.70-0.72) and the DNN model at 0.64 (95% CI 0.64-0.65). Due to the low prevalence of critical care outcomes, all models had high NPV with a 95% CI ranging from 0.98 to 0.99.

The number of the actual and predicted outcomes according to the level of KTAS is provided in Table 4. For the validation data set, the baseline model correctly identified 469 patients needing critical care in triage levels 1 and 2, which accounted for 65.0% of all critical care outcomes. However, it overtriaged 2296 patients in these high acuity categories. Undertriaging 35% of patients in need of critical care, the conventional model using a single variable of KTAS failed to predict all critical care outcomes (253 cases) for triage levels 3 to 5. Compared to the baseline model, the XGB model reduced false-positive cases from 2296 to 1533 in KTAS levels 1 and 2 and the false-negative cases from 253 to 80 in KTAS levels 3 to 5.

Table 2. Discrimination, reclassification, and calibration of critical care outcome prediction models for the validation cohort.

Model	Discrimination		Reclassification		Calibration
	AUROC[a] (95% CI)	*P* value[b]	NRI[c] (95% CI)	*P* value	H-L[d] test, *P* value
KTAS[e]	0.796 (0.781-0.811)	Reference	Reference	Reference	.80
XGB[f]	0.861 (0.848-0.874)	<.001	0.293 (0.219-0.366)	<.001	.24
DNN[g]	0.833 (0.819-0.848)	<.001	0.032 (0.024-0.041)	<.001	<.001

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

[b] *P* value for AUROC was calculated using DeLong's test.

[c] NRI: net reclassification index.

[d] H-L: Hosmer-Lemeshow test.

[e] KTAS: Korean Triage and Acute Scale.

[f] XGB: extreme gradient boosting.

[g] DNN: deep neural network.

Figure 2. Area under the receiver operating characteristic curve for validation data set. DNN: deep neural network; KTAS: Korean Triage and Acute Scale; XGB: extreme gradient boosting.

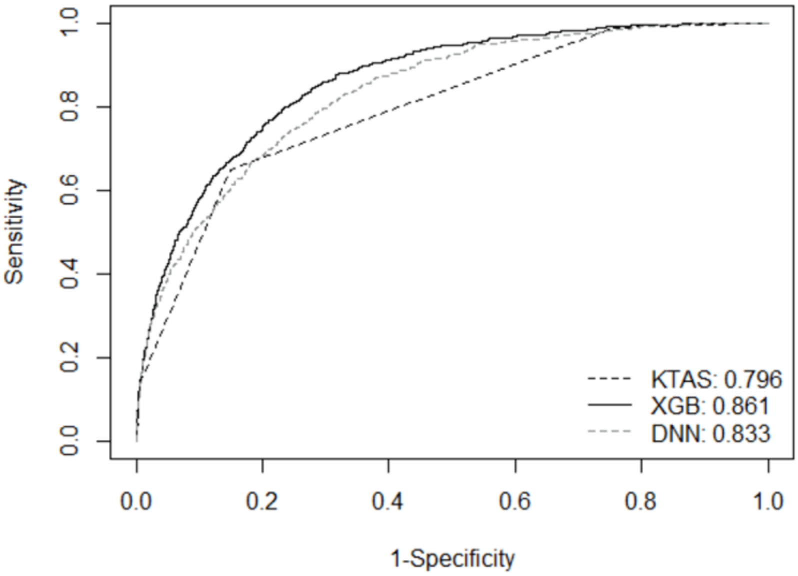


Figure 3. Calibration plot for validation data set. DNN: deep neural network; H-L test: Hosmer-Lemeshow test; KTAS: Korean Triage and Acute Scale; XGB: extreme gradient boosting. The observed probability of critical care with 95% CI is plotted against predicted probability by 10% interval. The diagonal line, which is represented as ideal, means perfect prediction. Point size indicates the relative number of observations in each bin.

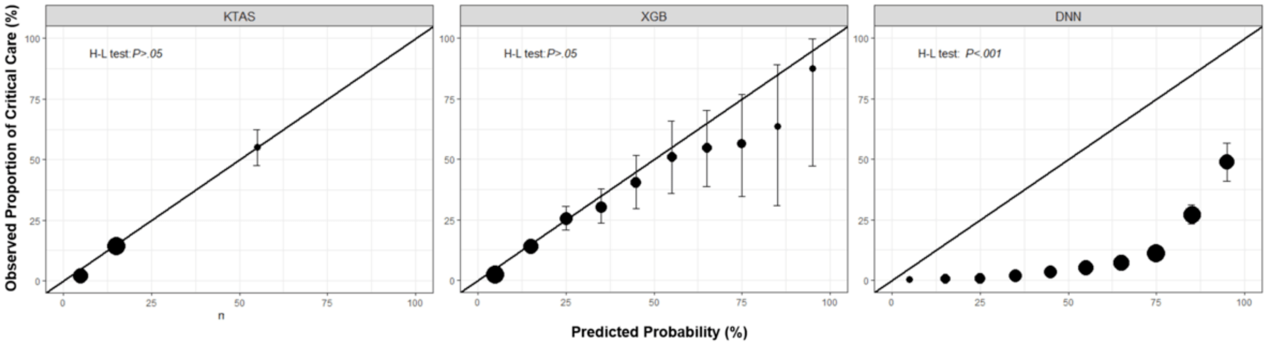


Table 3. Performance of critical care outcome prediction models in validation cohorts.

Model	Cutoff score	TP[a]	FP[b]	TN[c]	FN[d]	Sensitivity (95% CI)	Specificity (95% CI)	PPV[e] (95% CI)	NPV[f] (95% CI)
Baseline KTAS[g]	0.156[h]	469	2296	13,069	253	0.65 (0.61-0.68)	0.85 (0.84-0.86)	0.17 (0.16-0.18)	0.98 (0.98-0.98)
XGB[i]	0.036	616	4476	10,889	106	0.85 (0.83-0.88)	0.71 (0.70-0.72)	0.12 (0.11-0.13)	0.99 (0.99-0.99)
DNN[j]	0.444	614	5475	9890	108	0.85 (0.82-0.88)	0.64 (0.64-0.65)	0.10 (0.09-0.11)	0.99 (0.99-0.99)

[a] TP: true positive.

[b] FP: false positive.

[c] TN: true negative.

[d] FN: false negative.

[e] PPV: positive predictive values.

[f] NPV: negative predictive values.

[g] KTAS: Korean Triage and Acute Scale.

[h] Cutoff probability of 0.156 for the baseline model by logistic regression corresponds to KTAS score of 2.

[i] XGB: extreme gradient boosting.

[j] DNN: deep neural network.

Table 4. The performance comparison of prediction models in validation cohorts according to the level of KTAS.

KTAS[a] level	Actual critical care, n (%)	Baseline model				XGB[b] model			
		TP[c]	FP[d]	TN[e]	FN[f]	TP	FP	TN	FN
1: Resuscitation (n=178, 1.1%)	98 (13.6)	98	80	0	0	96	77	3	2
2: Emergent (n=2587, 16.1%)	371 (51.4)	371	2216	0	0	347	1456	760	24
3: Urgent (n=9559, 59.4%)	244 (33.8)	0	0	9315	244	170	2622	6693	74
4: Less urgent (n=3312, 20.6%)	9 (1.2)	0	0	3303	9	3	297	3006	6
5: Nonurgent (n=451, 2.8%)	0 (0.0)	0	0	451	0	0	24	427	0
Total (n=16,086, 100%)	722 (100)	469	2296	13,069	253	616	4476	10,889	106

[a] KTAS: Korean Triage and Acute Scale.

[b] XGB: extreme gradient boosting.

[c] TP: true positive.

[d] FP: false positive.

[e] TN: true negative.

[f] FN: false negative.

Variable Importance and Partial Dependence Plot

We computed permutation-based variable importance for the XGB and DNN model in Figure 4. The variable ranked as a top priority was chief complaints for the XGB model and EMS use for the DNN model. Despite the ranking difference in variable importance between the XGB and DNN models, variables higher in the list, including chief complaints, EMS use, age, AVPU, PR, and RR, were identical.

For the XGB model defined as the final prediction model, the relationship between each variable and the prediction outcome for the validation data set is illustrated in Figure 5. The PDP shows the marginal effect of a single variable on the prediction outcome. The value of the y-axis on PDP is the predicted probability for critical care. The nonlinear associations of all vital sign variables to critical outcome predictions were demonstrated. For age, RR, and SpO_2, we found the pattern of the critical care prediction in the XGB model, indicating the probability of being classified as patients in need of critical care increased with older age, higher RR, and lower SpO_2. For SBP, DBP, and PR, we observed a U-shaped relationship between each vital sign and the critical care prediction.

Figure 4. Feature importance. The time interval denotes the time between onset and ED arrival. AUROC: area under the receiver operating characteristic curve; AVPU: alert, verbal, painful, and unresponsive; BT: body temperature; DBP: diastolic blood pressure; DNN: deep neural network; ED: emergency department; EMS: emergency medical service; PR: pulse rate; RR: respiratory rate; SBP: systolic blood pressure; SpO_2: oxygen saturation; XGB: extreme gradient boosting.

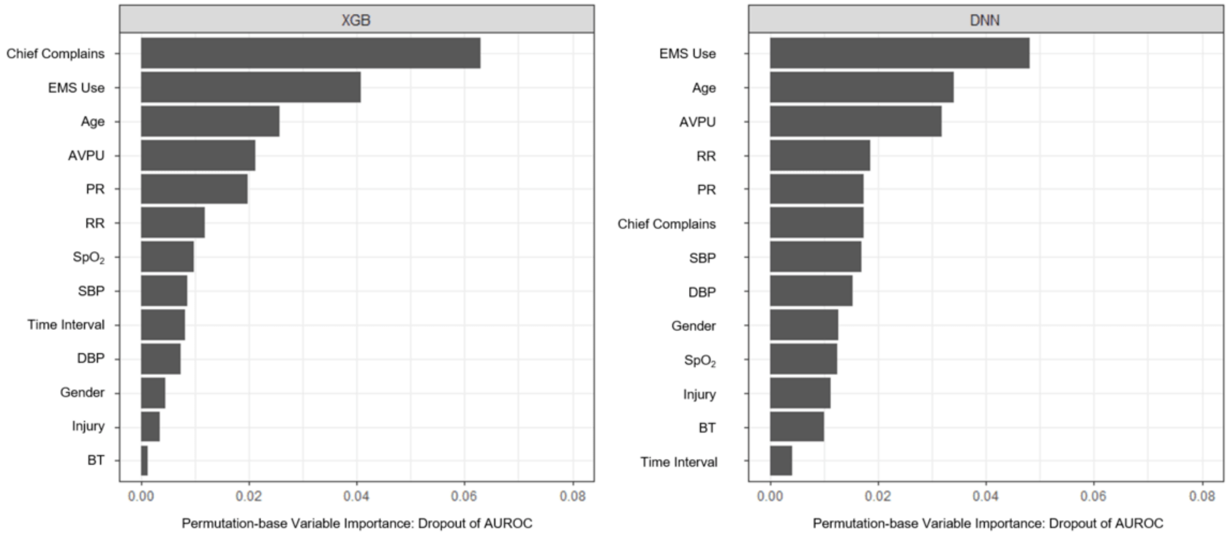
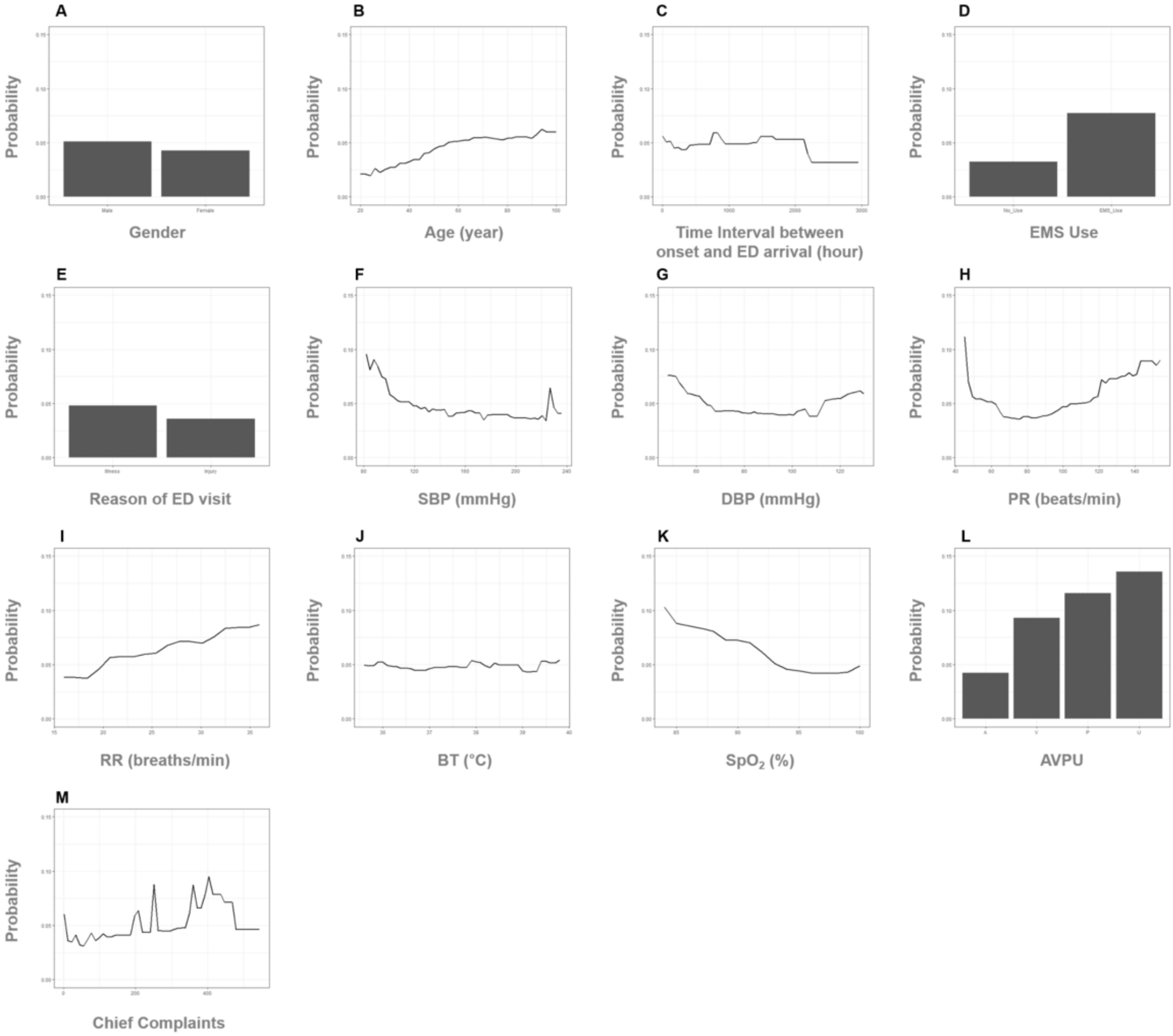


Figure 5. Partial dependence plot. A. gender, B. age, C. time interval between onset and ED arrival, D. EMS use, E. reason of ED visit, F. SBP, G. DBP, H. PR, I. RR, J. BT, K. SpO_2, L. AVPU, and M. chief complaints. The partial dependence plot shows the marginal effect of a single variable on the prediction outcome; the value of the y-axis is the predicted probability for critical care. AVPU: alert, verbal, pain, and unresponsive; BT: body temperature; DBP: diastolic blood pressure; ED: emergency department; PR: pulse rate; RR: respiratory rate; SBP: systolic blood pressure; SpO_2: oxygen saturation; XGB: extreme gradient boosting.



Discussion

Principal Findings

In this study, based on the data of 80,433 ED adult patients, we applied two modern machine learning approaches (ie, XGB and DNN) to the routinely collected triage information (age, gender, mode of ED arrival, the time interval between onset and ED arrival, reason of ED visit, chief complaints, six vital signs, and level of consciousness) for the critical care outcome prediction in ED. The prediction models demonstrated superior performance of discrimination from AUROC 0.833 to AUROC 0.861 for the validation cohort and net reclassification compared to the conventional baseline model using KTAS (AUROC 0.796). The XGB model showed better discriminating power (AUROC 0.861) than the DNN model. We revealed that the XGB model was well-calibrated in predicting critical care outcomes (Hosmer-Lemeshow test; *P*>.05).

The objective of this study was to accurately differentiate high-risk patients from the less urgent patients at the triage stage in the ED. Expedited evaluation and ED care of patients with critical illnesses are crucial for maximizing clinical outcomes, providing a strong rationale for their prediction at triage [7,42]. Previous studies have documented that current five-level triage systems (eg, ESI, MTS, and KTAS) have a suboptimal ability to identify patients at high risk, low inter-rater agreement, and high variability within the same triage level [4,6-10]. Hence, machine learning models incorporating variables of demographics, mode of ED arrival, chief complaints, and vital signs extracted from triage information have been investigated to support accurate and rapid decision-making of ED clinicians. This study extends the earlier research. The discriminative performance gains of the critical care outcome prediction were obtained from the XGB algorithm, which has the excellence to handle nonlinear interactions between variables and the prediction outcome.

In this study, a large number (85.5%) of the patients without a need of critical care were classified into KTAS levels 3 to 5 (83.2% of the entire population), while the majority (64.8%) of the critically ill patient group was assigned into KTAS level 1 and 2 (16.8 % of all patients). We demonstrated that the XGB model correctly detected critically ill patients who were undertriaged into lower-acuity KTAS levels 3 to 5 in the baseline model. The ability to reduce false-negative cases provides a strong rationale for adopting the machine learning algorithm model at ED triage, where the accurate and rapid identification of patients at high risk is a matter of the utmost importance. Furthermore, we observed that the XGB model reduced the number of false-positive cases that were overtriaged into high-acuity levels 1 to 2 in the baseline model, which may prevent excessive resource utilization in ED practices.

This research proved that the XGB model had agreement between the predicted probability and the observed proportion of critical care occurrences. The calibration plot in Figure 3 visualized how well the forecast probabilities from the XGB model were calibrated. Despite the importance of calibration in the prediction model to support clinician decision, systematic reviews have found that calibration is assessed far less than discrimination [24,25,27], which is problematic since poor calibration can make predictions misleading [24,26-28]. Machine learning algorithms are vulnerable to overfitting [24,33,43]. Due to overfitting, most machine learning algorithms, especially neural networks, are known to produce poor calibration when validated with new data [24,33,44,45]. However, XGB controls the model complexity by embedding a regularization term into the objective function to avoid overfitting [40,46,47]. Our findings suggest that the probabilities of the XGB model for predicting patients at high risk in ED were reliable.

Explaining the predictions of block-box machine learning has become highlighted. For the global interpretation of the model, we visualized the nonlinear relationship between a variable and outcome results in predicting critically ill patients using PDPs (Figure 5). The XGB algorithm interpreted that, on average, higher RR and lower SpO_2 are associated with a high probability of critical care outcomes, and there was a U-shaped relationship between SBP, DBP, and PR and the outcome results. The interpretation of the XGB model clearly reflected the characteristics of vital signs and was in line with medical knowledge. There are several interpretation techniques for global and local levels of machine learning interpretation. A future study of the multilevel interpretation of machine learning algorithm predictions is warranted.

Using triage information and the XGB algorithm, the artificial intelligent model for predicting patients at high risk in this study can be implemented in the ED setting without additional burden, which may support prompt and accurate clinician decision-making at the early stage of ED triage, leading to the improvement of patients' health outcomes and contributing to efficient ED resource allocation.

Limitations

This study has several limitations. First, we used the data from a single ED of a tertiary-care university hospital; therefore, external validation is needed for the generalization of the results. Second, this study did not address how the prediction model could be deployed into the clinical pathway; therefore, future studies applying the prediction model during triage are warranted.

Conclusions

This study demonstrated that using initial triage information routinely collected in the ED, the machine learning model improved the discrimination and net reclassification for predicting patients in need of critical care in ED compared to the conventional approach with KTAS. Moreover, we demonstrated that the XGB model was well-calibrated and interpreted nonlinear characteristics of vital sign predictors in line with medical knowledge.

Authors' Contributions

HY designed the study, developed and implemented modeling methods, analyzed modeling results, and drafted the original manuscript. JHP contributed to data interpretation and the final drafting of the manuscript. JC supervised the model development and evaluation and contributed to the final drafting of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Supplementary files including (1) variable description for clinical outcome prediction in emergency department, (2) hyperparameter optimization, and (3) comparison of baseline characteristics of study population between training and validation datasets.
[DOCX File , 30 KB - medinform_v9i9e30770_app1.docx]

References

1. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. PLoS One 2018;13(8):e0203316 [FREE Full text] [doi: 10.1371/journal.pone.0203316] [Medline: 30161242]

2. Zachariasse JM, van der Hagen V, Seiger N, Mackway-Jones K, van Veen M, Moll HA. Performance of triage systems in emergency care: a systematic review and meta-analysis. BMJ Open 2019 May 28;9(5):e026471 [FREE Full text] [doi: 10.1136/bmjopen-2018-026471] [Medline: 31142524]
3. Graham B, Bond R, Quinn M, Mulvenna M. Using Data Mining to Predict Hospital Admissions From the Emergency Department. IEEE Access 2018;6:10458-10469. [doi: 10.1109/access.2018.2808843]
4. Dugas AF, Kirsch TD, Toerper M, Korley F, Yenokyan G, France D, et al. An Electronic Emergency Triage System to Improve Patient Distribution by Critical Outcomes. J Emerg Med 2016 Jul;50(6):910-918. [doi: 10.1016/j.jemermed.2016.02.026] [Medline: 27133736]
5. Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JM. Clinical Decision Support Systems for Triage in the Emergency Department using Intelligent Systems: a Review. Artif Intell Med 2020 Jan;102:101762. [doi: 10.1016/j.artmed.2019.101762] [Medline: 31980099]
6. Fernandes M, Mendes R, Vieira SM, Leite F, Palos C, Johnson A, et al. Predicting Intensive Care Unit admission among patients presenting to the emergency department using machine learning and natural language processing. PLoS One 2020 Mar 3;15(3):e0229331 [FREE Full text] [doi: 10.1371/journal.pone.0229331] [Medline: 32126097]
7. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. Ann Emerg Med 2018 May;71(5):565-574.e2. [doi: 10.1016/j.annemergmed.2017.08.005] [Medline: 28888332]
8. 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]
9. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019 Mar 22;23(1):64 [FREE Full text] [doi: 10.1186/s13054-019-2351-7] [Medline: 30795786]
10. Goto T, Camargo CA, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 2018 Sep;36(9):1650-1654. [doi: 10.1016/j.ajem.2018.06.062] [Medline: 29970272]
11. Choi SW, Ko T, Hong KJ, Kim KH. Machine Learning-Based Prediction of Korean Triage and Acuity Scale Level in Emergency Department Patients. Healthc Inform Res 2019 Oct;25(4):305-312 [FREE Full text] [doi: 10.4258/hir.2019.25.4.305] [Medline: 31777674]
12. Kwon J, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS One 2018 Oct 15;13(10):e0205836 [FREE Full text] [doi: 10.1371/journal.pone.0205836] [Medline: 30321231]
13. Lee JH, Park YS, Park IC, Lee HS, Kim JH, Park JM, et al. Over-triage occurs when considering the patient's pain in Korean Triage and Acuity Scale (KTAS). PLoS One 2019 May 9;14(5):e0216519 [FREE Full text] [doi: 10.1371/journal.pone.0216519] [Medline: 31071132]
14. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng 2018 Oct 10;2(10):749-760 [FREE Full text] [doi: 10.1038/s41551-018-0304-0] [Medline: 31001455]
15. Elshawi R, Al-Mallah MH, Sakr S. On the interpretability of machine learning-based model for predicting hypertension. BMC Med Inform Decis Mak 2019 Jul 29;19(1):146 [FREE Full text] [doi: 10.1186/s12911-019-0874-0] [Medline: 31357998]
16. Mahmoudi E, Kamdar N, Kim N, Gonzales G, Singh K, Waljee AK. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. BMJ 2020 Apr 08;369:m958 [FREE Full text] [doi: 10.1136/bmj.m958] [Medline: 32269037]
17. Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. PLoS Med 2018 Nov 20;15(11):e1002695 [FREE Full text] [doi: 10.1371/journal.pmed.1002695] [Medline: 30458006]
18. Goto T, Camargo CA, Faridi MK, Freishtat RJ, Hasegawa K. Machine Learning-Based Prediction of Clinical Outcomes for Children During Emergency Department Triage. JAMA Netw Open 2019 Jan 04;2(1):e186937 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.6937] [Medline: 30646206]
19. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018 Jul 20;13(7):e0201016 [FREE Full text] [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
20. Zlotnik A, Alfaro MC, Pérez MCP, Gallardo-Antolín A, Martínez JMM. Building a Decision Support System for Inpatient Admission Prediction With the Manchester Triage System and Administrative Check-in Variables. Comput Inform Nurs 2016 May;34(5):224-230. [doi: 10.1097/CIN.0000000000000230] [Medline: 26974710]
21. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. JAMA 2017 Oct 10;318(14):1377-1384. [doi: 10.1001/jama.2017.12126] [Medline: 29049590]

22. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. Eur Heart J 2014 Aug 01;35(29):1925-1931 [FREE Full text] [doi: 10.1093/eurheartj/ehu207] [Medline: 24898551]
23. Steyerberg E, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010 Jan;21(1):128-138 [FREE Full text] [doi: 10.1097/EDE.0b013e3181c30fb2] [Medline: 20010215]
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group ‘Evaluating diagnostic testsprediction models’ of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med 2019 Dec 16;17(1):230 [FREE Full text] [doi: 10.1186/s12916-019-1466-7] [Medline: 31842878]
25. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019 Jun;110:12-22. [doi: 10.1016/j.jclinepi.2019.02.004] [Medline: 30763612]
26. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. Diagn Progn Res 2017 Dec 21;1(1):20 [FREE Full text] [doi: 10.1186/s41512-017-0021-2] [Medline: 31093549]
27. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol 2014 Mar 19;14(1):40 [FREE Full text] [doi: 10.1186/1471-2288-14-40] [Medline: 24645774]
28. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. Med Decis Making 2015 Feb;35(2):162-169. [doi: 10.1177/0272989X14547233] [Medline: 25155798]
29. Tonekaboni S. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR 106:359-380, 2019 2019;106:359-380 [FREE Full text]
30. Ogunleye A, Wang Q. XGBoost Model for Chronic Kidney Disease Diagnosis. IEEE/ACM Trans Comput Biol and Bioinf 2020 Nov 1;17(6):2131-2140. [doi: 10.1109/tcbb.2019.2911071]
31. Huang Z, Hu C, Chi C, Jiang Z, Tong Y, Zhao C. An Artificial Intelligence Model for Predicting 1-Year Survival of Bone Metastases in Non-Small-Cell Lung Cancer Patients Based on XGBoost Algorithm. Biomed Res Int 2020 Jun 28;2020:3462363-3462313 [FREE Full text] [doi: 10.1155/2020/3462363] [Medline: 32685470]
32. Spangler D, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. PLoS One 2019 Dec 13;14(12):e0226518 [FREE Full text] [doi: 10.1371/journal.pone.0226518] [Medline: 31834920]
33. Hou C, Zhong X, He P, Xu B, Diao S, Yi F, et al. Predicting Breast Cancer in Chinese Women Using Machine Learning Techniques: Algorithm Development. JMIR Med Inform 2020 Jul 08;8(6):e17364 [FREE Full text] [doi: 10.2196/17364] [Medline: 32510459]
34. Huang Y, Li W, Macheret F, Gabriel RA, Ohno-Machado L. A tutorial on calibration measurements and calibration models for clinical prediction models. J Am Med Inform Assoc 2020 Apr 01;27(4):621-633 [FREE Full text] [doi: 10.1093/jamia/ocz228] [Medline: 32106284]
35. Cava WL, Bauer C, Moore JH, Pendergrass SA. Interpretation of machine learning predictions for patient outcomes in electronic health records. AMIA Annu Symp Proc 2019;2019:572-581 [FREE Full text] [Medline: 32308851]
36. Zhou W, Wang Y, Gu X, Feng Z, Lee K, Peng Y, et al. Importance of general adiposity, visceral adiposity and vital signs in predicting blood biomarkers using machine learning. Int J Clin Pract 2021 Jan 26;75(1):e13664. [doi: 10.1111/ijcp.13664] [Medline: 32770817]
37. Muhlestein W, Akagi D, Kallos J, Morone P, Weaver K, Thompson R, et al. Using a Guided Machine Learning Ensemble Model to Predict Discharge Disposition following Meningioma Resection. J Neurol Surg B Skull Base 2018 May 08;79(2):123-130 [FREE Full text] [doi: 10.1055/s-0037-1604393] [Medline: 29868316]
38. Gómez-Ramírez J, Ávila-Villanueva M, Fernández-Blázquez MA. Selecting the most important self-assessed features for predicting conversion to mild cognitive impairment with random forest and permutation-based methods. Sci Rep 2020 Nov 26;10(1):20630 [FREE Full text] [doi: 10.1038/s41598-020-77296-4] [Medline: 33244011]
39. Delfin C, Krona H, Andiné P, Ryding E, Wallinius M, Hofvander B. Prediction of recidivism in a long-term follow-up of forensic psychiatric patients: Incremental effects of neuroimaging data. PLoS One 2019 May 16;14(5):e0217127 [FREE Full text] [doi: 10.1371/journal.pone.0217127] [Medline: 31095633]
40. Rzychoń M, Żogała A, Róg L. Experimental study and extreme gradient boosting (XGBoost) based prediction of caking ability of coal blends. Journal of Analytical and Applied Pyrolysis 2021 Jun;156:105020. [doi: 10.1016/j.jaap.2021.105020]
41. Roger E, Torlay L, Gardette J, Mosca C, Banjac S, Minotti L, et al. A machine learning approach to explore cognitive signatures in patients with temporo-mesial epilepsy. Neuropsychologia 2020 May;142:107455 [FREE Full text] [doi: 10.1016/j.neuropsychologia.2020.107455] [Medline: 32272118]
42. Elliott DJ, Williams KD, Wu P, Kher HV, Michalec B, Reinbold N, et al. An Interdepartmental Care Model to Expedite Admission from the Emergency Department to the Medical ICU. Jt Comm J Qual Patient Saf 2015 Dec;41(12):542-549. [doi: 10.1016/s1553-7250(15)41071-2] [Medline: 26567144]

43. Zhang Z, Ho KM, Hong Y. Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. Crit Care 2019 May 08;23(1):112 [FREE Full text] [doi: 10.1186/s13054-019-2411-z] [Medline: 30961662]
44. Kull M, Perello-Nieto M, Kängsepp M, Silva Filho T, Song H, Flach P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. In: Advances in Neural Information Processing Systems 32 (NeurIPS 2019). 2019 Presented at: Thirty-third Conference on Neural Information Processing Systems; December 8-14, 2019; Vancouver, Canada URL: https://proceedings.neurips.cc/paper/2019
45. Guo C. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning in Proceedings of Machine Learning Research 70: PMLR, 2017. 2017 Presented at: International Conference on Machine Learning; August 6-11, 2017; Sydney, Australia p. 1321-1330 URL: https://proceedings.mlr.press/v70/guo17a.html
46. Do DT, Le NQK. Using extreme gradient boosting to identify origin of replication in Saccharomyces cerevisiae via hybrid features. Genomics 2020 May;112(3):2445-2451. [doi: 10.1016/j.ygeno.2020.01.017] [Medline: 31987913]
47. Wu T, Chen H, Jhou M, Chen Y, Chang T, Lu C. Evaluating the Effect of Topical Atropine Use for Myopia Control on Intraocular Pressure by Using Machine Learning. J Clin Med 2020 Dec 30;10(1):111 [FREE Full text] [doi: 10.3390/jcm10010111] [Medline: 33396943]

Abbreviations

AUROC: area under the receiver operating characteristic
AVPU: alert, verbal, painful, and unresponsive
BT: body temperature
CTAS: Canadian Triage and Acuity Scale
DBP: diastolic blood pressure
DNN: deep neural network
ED: emergency department
ESI: emergency severity index
ICU: intensive care unit
KTAS: Korean Triage and Acute Scale
MTS: Manchester Triage System
NPV: negative predictive value
NRI: net reclassification index
PDP: partial dependence plot
PPV: positive predictive value
PR: pulse rate
RR: respiratory rate
SBP: systolic blood pressure
XGB: extreme gradient boosting

Edited by G Eysenbach; submitted 28.05.21; peer-reviewed by T Goto; comments to author 18.06.21; revised version received 27.06.21; accepted 27.07.21; published 20.09.21.

Please cite as:
Yun H, Choi J, Park JH
Prediction of Critical Care Outcome for Adult Patients Presenting to Emergency Department Using Initial Triage Information: An XGBoost Algorithm Analysis
JMIR Med Inform 2021;9(9):e30770
URL: https://medinform.jmir.org/2021/9/e30770
doi:10.2196/30770
PMID:34346889

Original Paper

Forecasting the Requirement for Nonelective Hospital Beds in the National Health Service of the United Kingdom: Model Development Study

Kanan Shah[1], BS; Akarsh Sharma[2], MSc; Chris Moulton[3], MBChB, FRCEM, MRCGP, FRCA; Simon Swift[4,5], MBBS, MRCS, MSc; Clifford Mann[6], MBBS; Simon Jones[7], PhD

[1]NYU Grossman School of Medicine, New York, NY, United States
[2]Icahn School of Medicine at Mount Sinai, New York, NY, United States
[3]The Royal Bolton Hospital, Bolton, United Kingdom
[4]Methods Analytics, London, United Kingdom
[5]University of Exeter Business School, Exeter, United Kingdom
[6]Taunton & Somerset NHS Foundation trust, Taunton, United Kingdom
[7]Division of Healthcare Delivery Science, Department of Population Health, NYU Grossman School of Medicine, New York, NY, United States

Corresponding Author:
Simon Jones, PhD
Division of Healthcare Delivery Science
Department of Population Health
NYU Grossman School of Medicine
227 E 30th St
New York, NY, 10016
United States
Phone: 1 646 501 2905
Email: simon.jones@nyulangone.org

Abstract

Background: Over the last decade, increasing numbers of emergency department attendances and an even greater increase in emergency admissions have placed severe strain on the bed capacity of the National Health Service (NHS) of the United Kingdom. The result has been overcrowded emergency departments with patients experiencing long wait times for admission to an appropriate hospital bed. Nevertheless, scheduling issues can still result in significant underutilization of bed capacity. Bed occupancy rates may not correlate well with bed availability. More accurate and reliable long-term prediction of bed requirements will help anticipate the future needs of a hospital's catchment population, thus resulting in greater efficiencies and better patient care.

Objective: This study aimed to evaluate widely used automated time-series forecasting techniques to predict short-term daily nonelective bed occupancy at all trusts in the NHS. These techniques were used to develop a simple yet accurate national health system–level forecasting framework that can be utilized at a low cost and by health care administrators who do not have statistical modeling expertise.

Methods: Bed occupancy models that accounted for patterns in occupancy were created for each trust in the NHS. Daily nonelective midnight trust occupancy data from April 2011 to March 2017 for 121 NHS trusts were utilized to generate these models. Forecasts were generated using the three most widely used automated forecasting techniques: exponential smoothing; Seasonal Autoregressive Integrated Moving Average; and Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components. The NHS Modernisation Agency's recommended forecasting method prior to 2020 was also replicated.

Results: The accuracy of the models varied on the basis of the season during which occupancy was forecasted. For the summer season, percent root-mean-square error values for each model remained relatively stable across the 6 forecasted weeks. However, only the trend and seasonal components model (median error=2.45% for 6 weeks) outperformed the NHS Modernisation Agency's recommended method (median error=2.63% for 6 weeks). In contrast, during the winter season, the percent root-mean-square error values increased as we forecasted further into the future. Exponential smoothing generated the most accurate forecasts (median error=4.91% over 4 weeks), but all models outperformed the NHS Modernisation Agency's recommended method prior to 2020 (median error=8.5% over 4 weeks).

Conclusions: It is possible to create automated models, similar to those recently published by the NHS, which can be used at a hospital level for a large national health care system to predict nonelective bed admissions and thus schedule elective procedures.

(JMIR Med Inform 2021;9(9):e21990) doi:10.2196/21990

KEYWORDS

bed occupancy; clinical decision-making; forecasting; health care delivery; models; time-series analysis

Introduction

Background and Rationale

Between 2011-2012 and 2019-2020, patient attendances at major (Type 1) emergency departments (EDs) in the National health Service (NHS) of the United Kingdom increased by approximately 20%. There was an even greater increase in the number of patients admitted to hospital from the ED during that time. Such admissions grew by more than one-third and now account for nearly three-fourth of all nonelective admitted patients (Figure 1).

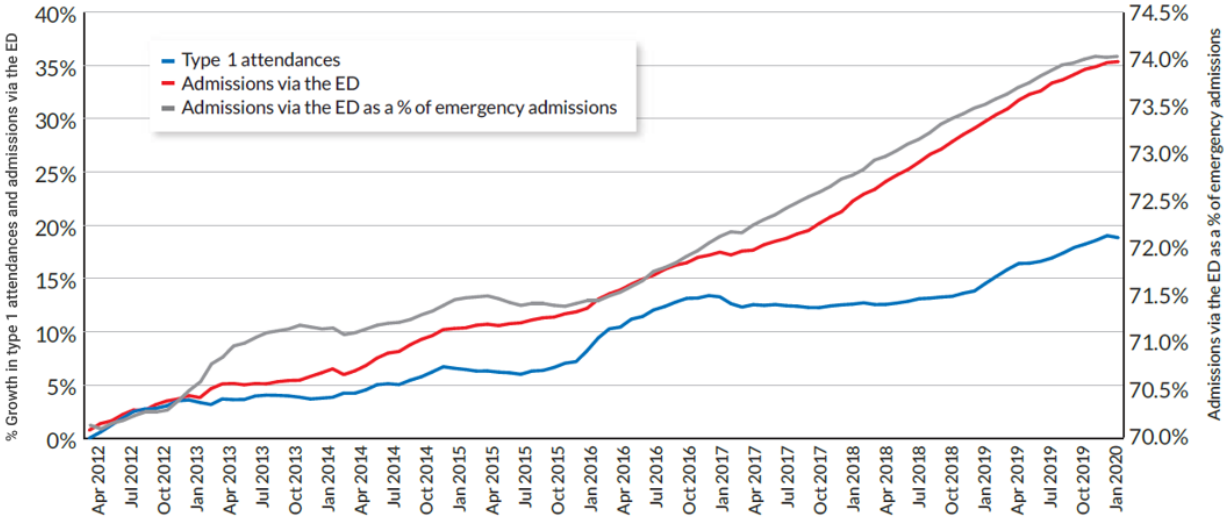
The resulting strain on the bed capacity of the NHS resulted in overcrowding of EDs and long wait times for patients before admission to an inpatient ward. By 2019-2020, over 3.2% of all patients in the ED remained in the ED for more than 12 hours from their time of arrival [1]. Such long delays are known to cause poor patient outcomes, including an increase in all-cause 30-day mortality [2].

For health care systems to meet the increasing needs of the populations they serve, as well as to provide better care, it is imperative to optimize the allocation of existing health care resources, including hospital beds. Health forecasting, a novel area of forecasting, can facilitate this by providing health service providers with the hospital bed occupancy forecasts that will allow them to minimize risks and manage demand [3].

Current models, including NHS-recommended systems, predict hospital bed utilization with a significant degree of error and with marked variability among different hospitals. More accurate and reliable long-term prediction of bed requirements will facilitate the anticipation of a local population's needs [4] with resulting gains in both efficiency and patient outcomes.

Many studies have attempted to conduct time-series analyses to forecast bed occupancy levels days or weeks in advance [5]. Most of these models use estimates of length of stay [6] or ED admissions [7-9]. We developed models using a more direct time-series–based approach, which utilizes historic admissions data without any identifying patient information, to model nonelective hospital bed requirements. Knowledge of future nonelective bed occupancy allows for proper scheduling of elective procedures to optimize both capacity and resource allocation [6]. In addition, no previous study, to our knowledge, has generated accurate models for an entire national health care system. In this study, we created a modeling framework that is automated and generalizable across the NHS of the United Kingdom. It can be used by administrators and key decision-makers, who have minimal knowledge of statistical techniques, to evaluate and respond more efficiently to clinical demand.

Figure 1. Growth in the admission rate of all nonelective admitted patients. ED: emergency department. Data source: NHS Hospital Episode Statistics for Accident & Emergency and for admitted patient care data, 2012-2020 [10].



Data source: AE HES & APC HES, 2012-2020

Objectives

We aimed to (1) test and compare a set of widely used automated time-series forecasting techniques to predict short-term (up to 42 days) nonelective bed occupancy on a daily basis and (2) develop a simple yet accurate system-level forecasting and modeling framework that could be used to predict emergency bed occupancy during different seasonal patterns of admission. A summary of the study rationale and objectives is provided in Textbox 1.

Textbox 1. Summary of the study objectives and rationale.

What is already known

Current statistical models that forecast bed occupancy days or weeks in advance across all trusts in the National Health Service (NHS) of the United Kingdom, including the previously recommended NHS method, predict hospital bed utilization with significant errors and variability among hospitals.

What this study adds

We created a modeling framework, following advanced forecasting techniques recommended by the NHS in January 2020, which generates similar forecasts of bed occupancy levels weeks in advance for all trusts in the NHS. In addition, because it is automated and generalizable, this model can be used by administrators and key decision-makers who have a minimal statistical background.

Methods

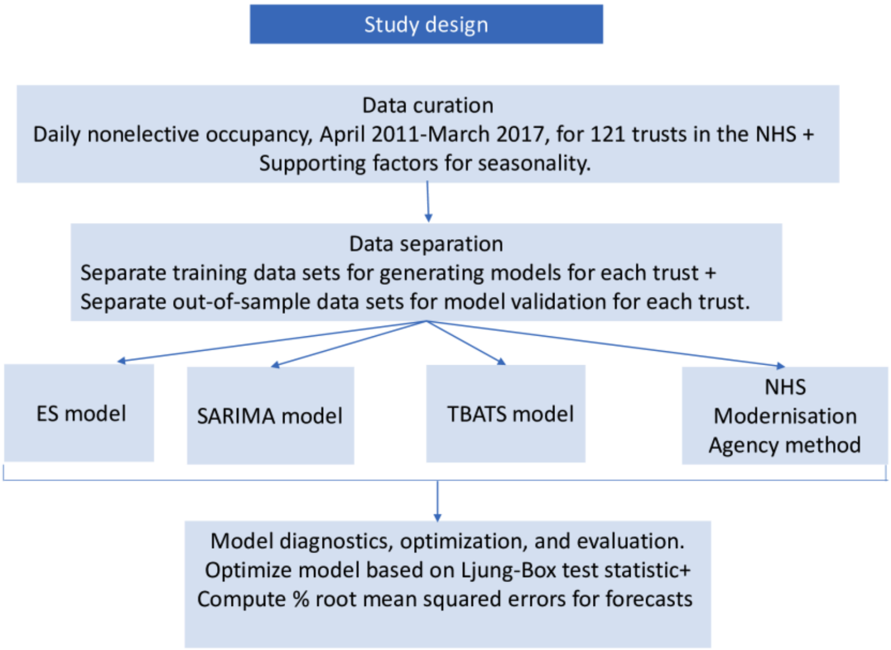
Data

Our data set contained daily nonelective midnight trust occupancy data from April 2011 to March 2017 for 121 NHS trusts located in each region of England. We acknowledge that this is not the same as peak occupancy, which often occurs in the middle of the working day. No personal information on patients or staff was provided. Since these data did not contain any identifying patient information, ethics approval from the institutional review board was not required. In addition, administrative data were utilized, and patients and the public were not involved in our study. We performed all analyses using RStudio (version 1.1.442, RStudio Inc). All generated forecasts accounted for patterns in occupancy or seasonality, resulting from the day of the week being forecasted. One forecast also factored in the day of the year, incidence of public holidays, and historical bed availability.

Study Design

Data preparation and analysis were performed in 4 steps for each forecasting technique employed (Figure 2).

Figure 2. Methodology employed to develop models and generate forecasted nonelective occupancy for each trust in the NHS. ES: exponential smoothing, NHS: National Health Service, SARIMA: Seasonal Autoregressive Integrated Moving Average, and TBATS: Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components.


Study design
Data curation
Daily nonelective occupancy, April 2011-March 2017, for 121 trusts in the NHS +
Supporting factors for seasonality.
Data separation
Separate training data sets for generating models for each trust +
Separate out-of-sample data sets for model validation for each trust.
ES model
SARIMA model
TBATS model
NHS
Modernisation
Agency method
Model diagnostics, optimization, and evaluation.
Optimize model based on Ljung-Box test statistic+
Compute % root mean squared errors for forecasts

Data Curation

We extracted daily nonelective occupancy data for 121 trusts in the NHS. To limit our data to general and acute bed occupancy, we excluded admissions in which the consultant's specialty was related to mental health, learning disabilities, or maternity. The first 10 days' and last 20 days' worth of occupancy data for each trust were removed from our data set to account for any edge effects creating inaccuracies in data reporting. The following supporting variables were also included, as they were likely to produce fluctuations or seasonality, in bed occupancy [11]: (1) day of the week, (2) day of the year, (3) public holidays, and (4) historical bed availability.

Separation of Data

Each trust's hospital occupancy data were divided into 2 seasonal data sets for summer and winter each, given that hospitals are more burdened during winter months, and this may introduce complications in the forecasting process. Each seasonal data set was further subsetted into a training data set that was used to develop the models and a validation data set that was used to cross-validate the models. The validation data sets for the summer season contained the last 6 weeks of occupancy data for mid-July to mid-August 2016 and those for winter contained the last 6 weeks of occupancy data for February to mid-March 2017. The derivation data sets contained all remaining data from April 2011 to the start date of the validation data sets and were used to develop the models.

Model Generation and Evaluation

We developed a set of models for each trust by using the three most widely used, automated forecasting techniques: exponential smoothing (ES); Seasonal Autoregressive Integrated Moving Average (SARIMA); and Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components (TBATS) (Table 1). Details regarding this model are provided in Multimedia Appendix 1. These models are in line with the NHS Modernisation Agency's newly released overview of advanced forecasting techniques that can be used to model NHS services [12]. We also replicated the NHS Modernisation Agency's previously recommended, albeit dated, forecasting method [13].

Therefore, we developed 4 models for each of the 121 trusts in the NHS. A program was developed in R to automate the entire process to minimize repetition and maximize efficiency. Two tests were applied for each forecast: the Ljung-Box test output (Multimedia Appendix 2) to measure for residual patterns of the models' errors that could be corrected for with additional modeling parameters, and root-mean-square error (RMSE) values from cross-validation. Absolute RMSE values were then converted to percentage errors, representing the average prediction error irrespective of sign (Multimedia Appendix 2). The numerator was the median RMSE of the forecasting method, and the denominator was the total number of general and acute beds in the hospital. The denominator, therefore, was the same for all methods. A comparative analysis of forecast accuracy was performed by comparing forecasted daily nonelective occupancy with actual nonelective occupancy in the out-of-sample data set for each week forecasted.

Forecasts were generated for the summer and winter. Given that summer school holidays in the United Kingdom usually occur from late July until early September and NHS data for England suggest that winter pressures mostly last from early January until the end of March, we forecasted the time periods within those timeframes. Summer forecasts were obtained from July to mid-August 2016 and winter forecasts were obtained from mid-February to mid-March 2017. Forecasts were compared to each other as well as to those derived from the NHS Modernisation Agency's recommended method.

We did not generate models using Prophet and artificial neural networks, which are 2 additional models recommended in the Modernisation Agency's overview, because these cannot be automated and applied across multiple trusts [12].

Table 1. Descriptive characteristics of automated time-series forecasting techniques used.

Characteristics	Models generated			
	Exponential smoothing	Seasonal Autoregressive Integrated Moving Average	Trend and Seasonal Components	Method recommended by the National Health Service Modernisation Agency
Statistical methods employed	Exponential weighted sum of previous observations	Combines auto-regression and moving average models	State space reconstruction	Forecast is the mean value of the past 6 weeks' bed occupancy for the day of the week being forecasted
Seasonality taken into account	Weekly	Weekly, yearly, monthly, public holidays, and historical bed availability	Weekly	Weekly
Additional modifications performed	LThe Ljung-Box test provided information on residual patterns of error; this was addressed by creating a model of the residuals of the forecast	Suspected model may be more accurate for trusts that do not approach maximum occupancy (occupancy=<95%); percent occupancy was introduced as a seasonal component	None	None

Results

A total of 484 models (n=4 per trust) were automatically developed using our modeling framework (Figure 3). Our Ljung-Box tests validated that autocorrelation is minimal, thus validating our choice of model.

The accuracy of our models varied on the basis of the season during which we forecasted occupancy. Percent RMSE values for each model remained relatively stable across the 6 weeks forecasted in the summer, indicating that the summer period is predictable (Figure 4). In addition, only our TBATS model (median error=2.45% for 6 weeks) outperformed the NHS Modernisation Agency’s recommended method (median error=2.63% for 6 weeks). TBATS yielded a median error of 1.98% for the first forecasted week and 3.01% for the sixth, while the NHS Modernisation Agency’s recommended method yielded a median error of 2.32% for the first forecasted week and 3.17% for the sixth.

In contrast, percent RMSE values increased as we forecasted further into the future during winter (Figure 5). Therefore, our study suggests that we are only able to generate relatively accurate forecasts 4 weeks into the future during winter. Significant weather events and disease outbreaks may contribute to this unpredictability. However, as current weather forecasting methods are unable to predict significant events accurately beyond 10 days, accounting for weather beyond this is impractical. ES performed the best (median error=4.91% over 4 weeks), but all models outperformed the NHS Modernisation Agency’s recommended method (median error=8.5% over 4 weeks). ES yielded a median error of 2.17% for the first forecasted week and 9.38% for the fourth, while the NHS Modernisation Agency’s recommended method yielded a median error of 5.12% for the first forecasted week and 13.62% for the sixth.

Five or fewer trusts failed to pass the Ljung-Box test of autocorrelation for the TBATS and SARIMA models, which suggested that the models could not be improved much further for accuracy. However, as a large proportion of trusts failed to pass for the ES model (40% for summer forecasts and 42% for winter forecasts), we developed a TBATS model to forecast the residuals of our predictions and incorporated these forecasted residuals into our original model. This modification, however, did not significantly improve forecast accuracy. We also suspected that forecasts may be more accurate for trusts that do not reach their maximum bed availability (less than 95% of total beds are occupied). Therefore, we subsetted trusts on the basis of maximum bed availability and generated separate SARIMA forecasts for each group. This increased each forecast accuracy for the winter but did not affect forecasts much for the summer.

Figure 3. Sample time series and TBATS forecasts generated for summer and winter for two trusts: Gateshead Health (A-C) and Mid Essex Hospital Services (D-F). (A) and (D) show plots of nonelective bed occupancy through the time period in the data set. In (B)-(C) and (E)-(F), the black lines represent the training data sets, the red lines represent the out-of-sample datasets, and the blue lines represent the occupancy forecasted by TBATS. TBATS: Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components.

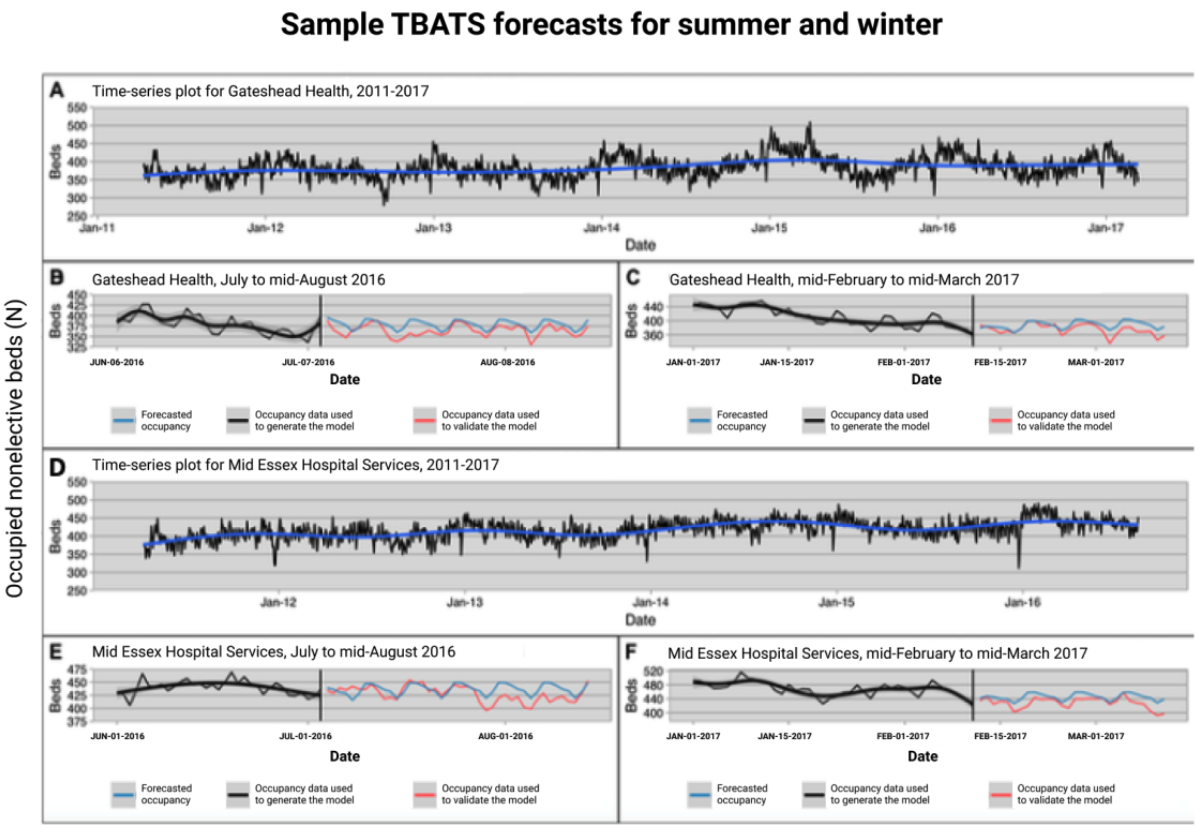


Figure 4. Distribution of error values for summer forecasts displayed by model (A) and by week (B). Outliers have been suppressed in (B) for better visualization of error spread. ES: exponential smoothing, NHS: National Health Service, RMS: root-mean-square, SARIMA: Seasonal Autoregressive Integrated Moving Average, TBATS: Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components.

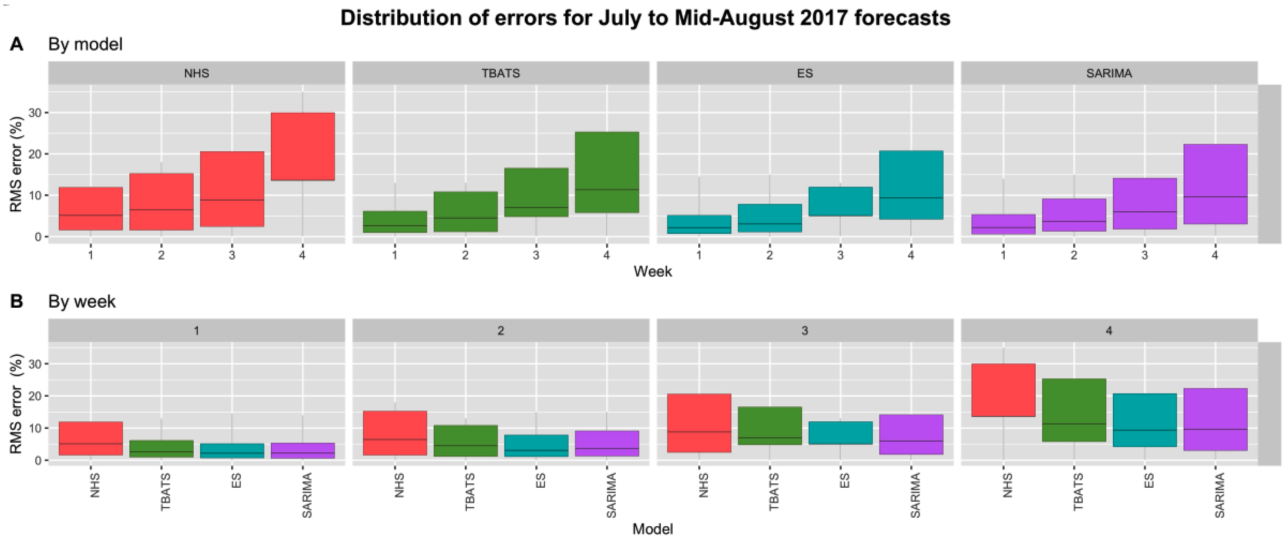
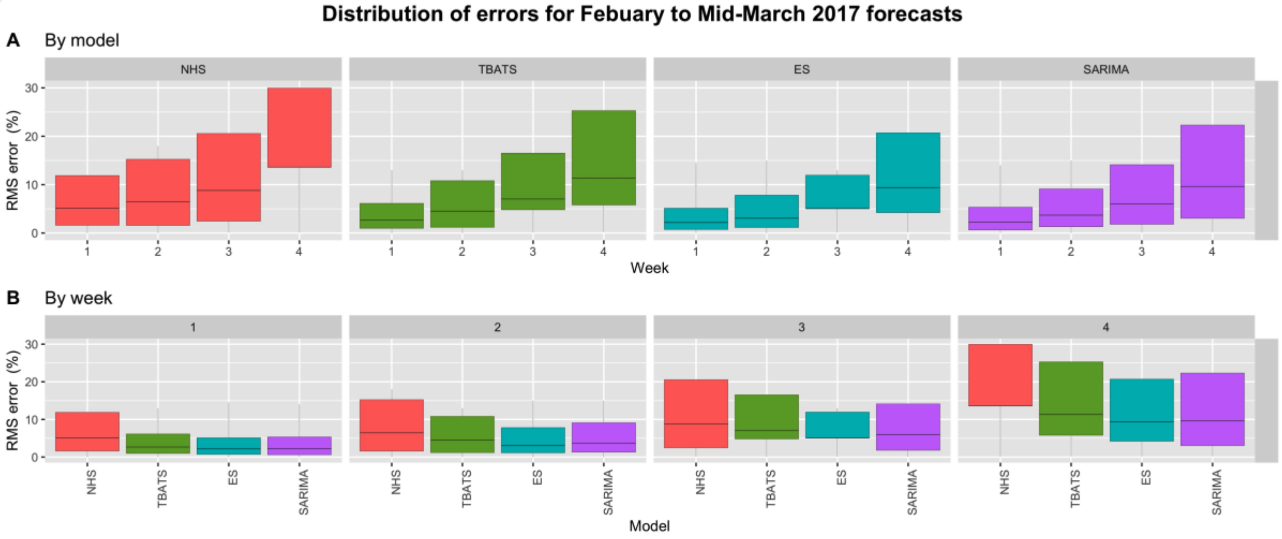


Figure 5. Distribution of error values for winter forecasts displayed by model (A) and by week (B). Outliers have been suppressed in (B) for better visualization of error spread. ES: exponential smoothing, NHS: National Health Service, RMS: root-mean-square, SARIMA: Seasonal Autoregressive Integrated Moving Average, TBATS: Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components.



Discussion

Principal Findings

Our results show that it is possible to create automated models to predict nonelective bed admissions with a higher degree of accuracy and reliability than the method previously recommended by the NHS Modernisation Agency.

Utilization of the NHS-recommended method has led to the lack of capacity and procedures being canceled at the last minute (Figure 3). Although some of these problems are inevitable in such a large and diffusely managed system, the frequency of such cancellations can be reduced with improved forecasting methods such as the one described in this study.

Other groups that focused on hospitals in England have generated forecasting models for either a single site or a group of trusts, or for specific hospital services such as the emergency department, which is more easily predictable [14]. However, to our knowledge, none of them have utilized the data of the entire NHS to generate more accurate forecasting models [14]. Although individual, site-specific models have the potential to outperform national models, we believe that the majority of hospitals do not have the resources, time, or expertise required to generate their own predictive methodology. Therefore, a more universal and easily implemented—albeit slightly less accurate—modeling framework is preferable. Moreover, our models are automated and require minimal effort for consistent execution.

Even with a simplified approach and appropriate end-user education, several barriers to implementation could limit the use of the developed national forecasting models. In the NHS system, staffing rotas lack the flexibility required to reduce staff at short notice. While an incorrect forecast predicting an increased demand would result in financial losses, an erroneous prediction of a reduction could lead to adverse clinical events. Therefore, caution dictates that users are more likely to respond to forecasts of increased demand rather than those predicting the reverse. Routine forecasting would therefore be likely to increase costs, at least in the initial phase.

If such modeling frameworks are to be incorporated into policy, it is essential to consider whether effective implementation is possible. A recent study of ED escalation plans [15] to support patient care in times of increased demand reported that there can be a significant gap between managerial intentions and actual implementation.

Limitations

Potential bias may have arisen from inaccurate data collection and reporting at a local level. In addition, our occupancy data were collected as a midnight census rather than during the day, when hospital occupancy peaks. This may limit the applicability of the model. Although our models took into account various temporal models, we did not explicitly consider meteorological conditions such as weather or air pollution—factors that could have an impact on predictive accuracy [16,17].

Another limitation of our study is that we were not able to model demand for outpatient services or elective procedures, which may have a significant impact on the availability of inpatient resources, including health care professionals themselves [18]. Nevertheless, this is an area that clinicians and managers can control to a large extent [13]. In addition, there is a need for caution when predicting the real-world implications of total bed utilization from models in which maximum capacity is approached, as small random effects may have unpredictable consequences.

As these models require automation, we have not utilized the most advanced predictive techniques available. The self-exciting threshold autoregressive model and artificial neural networks would be likely to produce more accurate predications once fully optimized. We explored this model but ultimately rejected it because of the high degree of personalization that the self-exciting threshold autoregressive system required for correct usage. This would thus be impractical for hospital staff with limited statistical knowledge—or the time to acquire it—for effective implementation.

Conclusions

There is no sign of an imminent reduction in the demand for all hospital services. Therefore, improvements in the efficiency of health care resource utilization are of paramount importance. We believe that, to our knowledge, this is the first study to generate accurate forecasting models for an entire health care system. In addition, our models are automated and require minimal effort to execute consistently and accurately; thus, they are in parallel with the NHS Modernisation Agency's latest guidelines on advanced forecasting techniques [12]. With increased predictive accuracy of nonelective bed occupancy, more reliable elective procedure schedules can be produced by hospital managers. This increased efficiency should lead to better care for patients, together with a more consistent workflow pattern for health care staff. We believe that a similar methodology can be applied to hospital systems other than the NHS, in other countries including the United States, and we hope to apply these models more widely in the future.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details regarding the forecasting techniques used in developing the current prediction model.

[DOCX File , 13 KB - medinform_v9i9e21990_app1.docx]

Multimedia Appendix 2

Supplementary tables.

[DOCX File , 284 KB - medinform_v9i9e21990_app2.docx]

References

1. Hospital Episode Statistics Data Dictionary. NHS Digital. URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics/hospital-episode-statistics-data-dictionary [accessed 2021-07-05]
2. Jones SM, Swift S, Moulton C, Molyneux P, Black S, Mason N, et al. The association between delays to patient admission from the Emergency Department and all-cause 30-day mortality. Emergency Medicine Journal 2021 (forthcoming).
3. Soyiri IN, Reidpath DD. An overview of health forecasting. Environ Health Prev Med 2013 Jan;18(1):1-9 [FREE Full text] [doi: 10.1007/s12199-012-0294-6] [Medline: 22949173]
4. Ordu M, Demir E, Tofallis C. A comprehensive modelling framework to forecast the demand for all hospital services. Int J Health Plann Manage 2019 Apr;34(2):e1257-e1271. [doi: 10.1002/hpm.2771] [Medline: 30901132]
5. Mackay M, Lee M. Choice of models for the analysis and forecasting of hospital beds. Health Care Manag Sci 2005 Aug;8(3):221-230. [doi: 10.1007/s10729-005-2013-y] [Medline: 16134435]

6. Kutafina E, Bechtold I, Kabino K, Jonas SM. Recursive neural networks in hospital bed occupancy forecasting. BMC Med Inform Decis Mak 2019 Mar 07;19(1):39 [FREE Full text] [doi: 10.1186/s12911-019-0776-1] [Medline: 30845940]
7. Hoot N, Epstein S, Allen T, Jones SS, Baumlin KM, Chawla N, et al. Forecasting emergency department crowding: an external, multicenter evaluation. Ann Emerg Med 2009 Oct;54(4):514-522.e19 [FREE Full text] [doi: 10.1016/j.annemergmed.2009.06.006] [Medline: 19716629]
8. Juang WC, Huang SJ, Huang FD, Cheng PW, Wann SR. Application of time series analysis in modelling and forecasting emergency department visits in a medical centre in Southern Taiwan. BMJ Open 2017 Dec 01;7(11):e018628 [FREE Full text] [doi: 10.1136/bmjopen-2017-018628] [Medline: 29196487]
9. Jones SA, Joy MP, Pearson J. Forecasting demand of emergency care. Health Care Manag Sci 2002 Nov;5(4):297-305. [doi: 10.1023/a:1020390425029] [Medline: 12437279]
10. Hospital Episode Statistics (HES). NHS Digital. URL: https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics [accessed 2021-09-29]
11. Rotstein Z, Wilf-Miron R, Lavi B, Shahar A, Gabbay U, Noy S. The dynamics of patient visits to a public hospital ED: a statistical model. Am J Emerg Med 1997 Oct;15(6):596-599. [doi: 10.1016/s0735-6757(97)90166-2] [Medline: 9337370]
12. NHS England and NHS Improvement. Advanced forecasting technique:. NHS. London: NHS England; 2020 Jan. URL: https://www.england.nhs.uk/wp-content/uploads/2020/01/advanced-forecasting-techniques.pdf [accessed 2021-08-31]
13. Proudlove NC, Black S, Fletcher A. OR and the challenge to improve the NHS: modelling for insight and improvement in in-patient flows. Journal of the Operational Research Society 2017 Dec 21;58(2):145-158. [doi: 10.1057/palgrave.jors.2602252]
14. Champion R, Kinsman LD, Lee GA, Masman KA, May EA, Mills TM, et al. Forecasting emergency department presentations. Aust Health Rev 2007 Feb;31(1):83-90. [doi: 10.1071/ah070083] [Medline: 17266491]
15. Back J, Ross A, Duncan M, Jaye P, Henderson K, Anderson J. Emergency Department Escalation in Theory and Practice: A Mixed-Methods Study Using a Model of Organizational Resilience. Ann Emerg Med 2017 Nov;70(5):659-671 [FREE Full text] [doi: 10.1016/j.annemergmed.2017.04.032] [Medline: 28662909]
16. Jilani T, Housley G, Figueredo G, Tang PS, Hatton J, Shaw D. Short and Long term predictions of Hospital emergency department attendances. Int J Med Inform 2019 Sep;129:167-174. [doi: 10.1016/j.ijmedinf.2019.05.011] [Medline: 31445251]
17. Poon CM, Wong ELY, Chau PYK, Yau SY, Yeoh EK. Management decision of hospital surge: assessing seasonal upsurge in inpatient medical bed occupancy rate among public acute hospitals in Hong Kong. QJM 2019 Jan 01;112(1):11-16. [doi: 10.1093/qjmed/hcy217] [Medline: 30295857]
18. Luo L, Luo L, Zhang X, He X. Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA and SES models. BMC Health Serv Res 2017 Jul 10;17(1):469 [FREE Full text] [doi: 10.1186/s12913-017-2407-9] [Medline: 28693579]

Abbreviations

ED: emergency department
ES: exponential smoothing
NHS: National Health Service
RMSE: root-mean-square error
SARIMA: Seasonal Autoregressive Integrated Moving Average
TBATS: Trigonometric, Box-Cox transform, autoregressive moving average errors, and Trend and Seasonal components

Edited by G Eysenbach; submitted 30.06.20; peer-reviewed by L Genaro, M Adly, A Adly, A Adly, A Azzam; comments to author 06.11.20; revised version received 15.05.21; accepted 03.06.21; published 30.09.21.

Please cite as:
Shah K, Sharma A, Moulton C, Swift S, Mann C, Jones S
Forecasting the Requirement for Nonelective Hospital Beds in the National Health Service of the United Kingdom: Model Development Study
JMIR Med Inform 2021;9(9):e21990
URL: https://medinform.jmir.org/2021/9/e21990
doi:10.2196/21990
PMID:

Original Paper

Measuring Collaboration Through Concurrent Electronic Health Record Usage: Network Analysis Study

Patrick Li[1], BSc; Bob Chen[2], BSc; Evan Rhodes[3], BSc; Jason Slagle[3], PhD; Mhd Wael Alrifai[4,5], MD; Daniel France[3], PhD; You Chen[5,6], PhD

[1]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, United States
[2]Epithelial Biology Center, Vanderbilt University Medical Center, Nashville, TN, United States
[3]Department of Anesthesiology, Vanderbilt University Medical Center, Nashville, TN, United States
[4]Department of Pediatric, Vanderbilt University Medical Center, Nashville, TN, United States
[5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[6]Department of Computer Science, Vanderbilt University, Nashville, TN, United States

Corresponding Author:
You Chen, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave
Nashville, TN
United States
Phone: 1 6153431939
Email: you.chen@vanderbilt.edu

Abstract

Background: Collaboration is vital within health care institutions, and it allows for the effective use of collective health care worker (HCW) expertise. Human-computer interactions involving electronic health records (EHRs) have become pervasive and act as an avenue for quantifying these collaborations using statistical and network analysis methods.

Objective: We aimed to measure HCW collaboration and its characteristics by analyzing concurrent EHR usage.

Methods: By extracting concurrent EHR usage events from audit log data, we defined concurrent sessions. For each HCW, we established a metric called concurrent intensity, which was the proportion of EHR activities in concurrent sessions over all EHR activities. Statistical models were used to test the differences in the concurrent intensity between HCWs. For each patient visit, starting from admission to discharge, we measured concurrent EHR usage across all HCWs, which we called temporal patterns. Again, we applied statistical models to test the differences in temporal patterns of the admission, discharge, and intermediate days of hospital stay between weekdays and weekends. Network analysis was leveraged to measure collaborative relationships among HCWs. We surveyed experts to determine if they could distinguish collaborative relationships between high and low likelihood categories derived from concurrent EHR usage. Clustering was used to aggregate concurrent activities to describe concurrent sessions. We gathered 4 months of EHR audit log data from a large academic medical center's neonatal intensive care unit (NICU) to validate the effectiveness of our framework.

Results: There was a significant difference (*P*<.001) in the concurrent intensity (proportion of concurrent activities: ranging from mean 0.07, 95% CI 0.06-0.08, to mean 0.36, 95% CI 0.18-0.54; proportion of time spent on concurrent activities: ranging from mean 0.32, 95% CI 0.20-0.44, to mean 0.76, 95% CI 0.51-1.00) between the top 13 HCW specialties who had the largest amount of time spent in EHRs. Temporal patterns between weekday and weekend periods were significantly different on admission (number of concurrent intervals per hour: 11.60 vs 0.54; *P*<.001) and discharge days (4.72 vs 1.54; *P*<.001), but not during intermediate days of hospital stay. Neonatal nurses, fellows, frontline providers, neonatologists, consultants, respiratory therapists, and ancillary and support staff had collaborative relationships. NICU professionals could distinguish high likelihood collaborative relationships from low ones at significant rates (3.54, 95% CI 3.31-4.37 vs 2.64, 95% CI 2.46-3.29; *P*<.001). We identified 50 clusters of concurrent activities. Over 87% of concurrent sessions could be described by a single cluster, with the remaining 13% of sessions comprising multiple clusters.

Conclusions: Leveraging concurrent EHR usage workflow through audit logs to analyze HCW collaboration may improve our understanding of collaborative patient care. HCW collaboration using EHRs could potentially influence the quality of patient care, discharge timeliness, and clinician workload, stress, or burnout.

(JMIR Med Inform 2021;9(9):e28998) doi:10.2196/28998

KEYWORDS

collaboration; electronic health records; audit logs; health care workers; neonatal intensive care unit; network analysis; clustering; visualization; concurrent interaction; human-computer interaction; survey instrument; informatics framework; secondary data analysis

Introduction

The measurement of coordinated collaboration in health care systems has proven to be important for providing better quality care [1-9]. Numerous studies have correlated collaboration with quality of care [1-3], patient safety [4-6], and clinical outcomes [7,8]. No universal guidelines exist to study collaboration in health care organizations (HCOs). Existing studies approached collaboration by relying on surveys, written reports, and interviews as a basis for gauging collaboration [1-9]. Further, they examined communication, teamwork, and problem-solving in HCOs, noting that interprofessional team functions are often suboptimal [3,7]. In addition, these studies identified barriers to successful interprofessional collaboration, including power dynamics, poor communication patterns, and incomplete understanding of roles and responsibilities [1-9]. However, existing studies seldom examine collaborative activities in the context of electronic health record (EHR) system usage. EHR systems provide a virtual environment for a diverse collection of health care workers (HCWs) to exchange accurate, detailed, and timely information electronically [10-12].

As EHRs have grown in adoption, the proportion of collaboration among HCWs involving EHR systems has increased as well [13-15]. For instance, a respiratory therapist noted, in an EHR, that a patient had an increased need for oxygen. At the same time, a nurse documented the same patient’s vital signs and noted the presentation of tachypnea. Next, an attending physician reviewed the vitals and respiratory rate, and prescribed the patient a diuretic [16]. Here, three HCWs experienced latent (inexplicit) collaboration through the EHR system that may not have been flagged by HCOs. HCWs may spend a considerable amount of time in latent collaborations in caring for patients through EHR systems [17-19]. The relationships among latent collaborations, care quality, and patient safety, however, have been understudied due to a lack of metrics or concepts describing collaboration of this nature.

Highly granular and widely available EHR audit logs document HCW activities occurring within EHRs [20-23] and can be used to model latent collaboration among HCWs and the respective interactions between HCWs and EHR systems [13,16,24-29]. Typically, each event documented in an audit log includes a timestamp, the type of action involved, the involved HCW and patient IDs, and further metadata, such as HCW specialties, patient demographics, and health conditions [13,16,20-29]. EHR audit logs have been widely used to measure health care organizational structures [20,25], clinical workflows [20,30,31], trauma care team structures [13,20,26], and intensive care unit care structures [16,27-29]. Existing studies have investigated audit log data at a coarse-grained level to build connections between HCWs [13,16,20-31], and thus, much of the contextual information (eg, HCW-EHR system interactions) is lost. For instance, coarse-grained latent interactions between HCWs have previously been defined by shared interactions with the same patients on the same day or during the same patient encounter [26-29]. We demonstrate that audit logs enable the study of latent collaborative activities at a highly granular level.

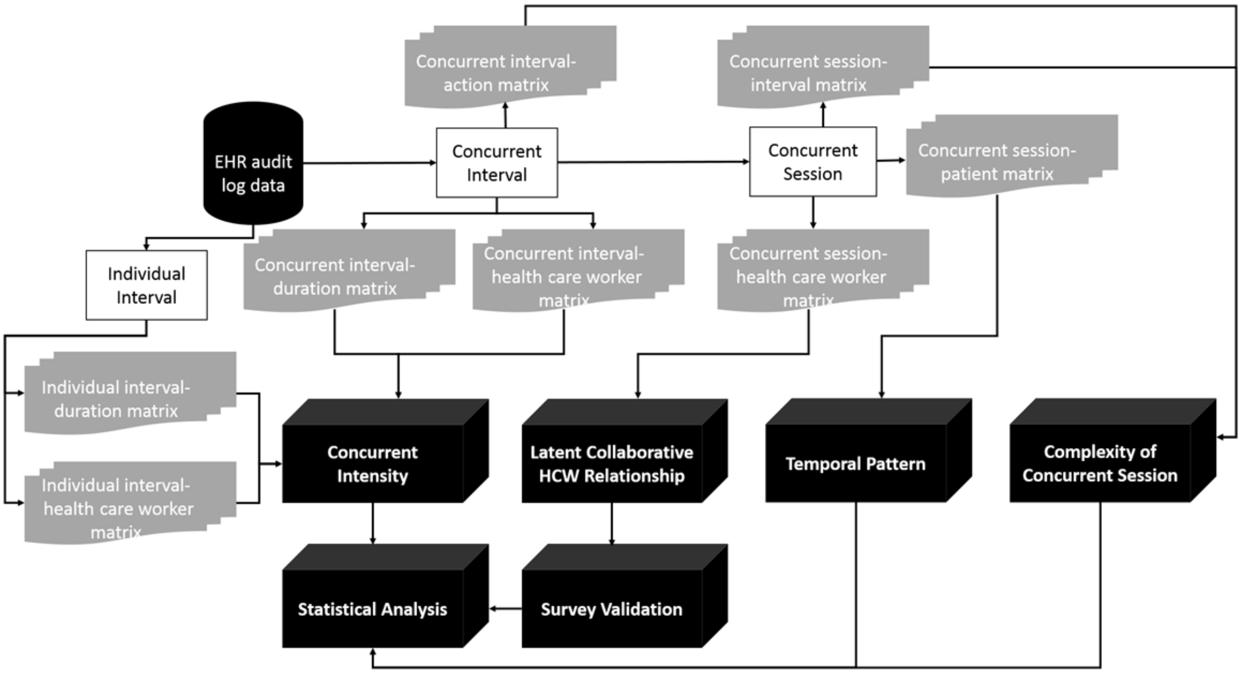
In this study, we propose a robust framework for the investigation of latent collaboration through concurrent EHR utilization. Using this framework, we describe a case study showcasing its usage in the neonatal intensive care unit (NICU) of a large academic medical center consisting of neonatologists, neonatal fellows, neonatal frontline providers, neonatal nurses, respiratory therapists, consultants, ancillary staff, and support staff. In the NICU, the density of audit logs per patient episode is very high, and it is an ideal environment for investigating latent collaboration [16,28].

Methods

Overview

In this section, we describe how we defined and calculated individual intervals, concurrent intervals, and concurrent sessions from the audit log data. We defined the core components of our proposed framework for measuring latent collaboration and its characteristics, via audit log data, which involve concurrent intensity (proportion of concurrent intervals and time spent on those intervals), latent collaborative HCW relationships, temporal patterns (weekday vs weekend or admission vs discharge temporal trends of concurrent EHR usage), and the complexity of concurrent sessions. Figure 1 shows the workflow of learning latent collaboration and its characteristics from the audit log data.

Figure 1. A workflow diagram showing our framework on learning concurrent intensity, latent collaborative HCW relationships, temporal trends of concurrent EHR usage, and concurrent session complexity from EHR audit log data. EHR: electronic health record; HCW: health care worker.


Concurrent interval-action matrix
Concurrent session-interval matrix
EHR audit log data
Concurrent Interval
Concurrent Session
Concurrent session-patient matrix
Individual Interval
Concurrent interval-duration matrix
Concurrent interval-health care worker matrix
Concurrent session-health care worker matrix
Individual interval-duration matrix
Individual interval-health care worker matrix
Concurrent Intensity
Latent Collaborative HCW Relationship
Temporal Pattern
Complexity of Concurrent Session
Statistical Analysis
Survey Validation

Events in EHR Audit Logs

An event is a single row of an audit log entry containing the HCW ID, patient ID, action ID, and time stamp. Thus, an event describes an action that an HCW performed on an EHR of a patient at a specific time. The action ID corresponds to the type of action performed, such as typing a progress note, accessing patient demographics, refilling medications, reviewing cholesterol test results, and so on. Table 1 shows a list of events performed by two HCWs (anonymized IDs A and B) on EHRs of two patients (anonymized IDs 1 and 2). These events are retrieved from EPIC EHR audit logs. Further definitions of the events can be found at Epic's EHR UserWeb [32].

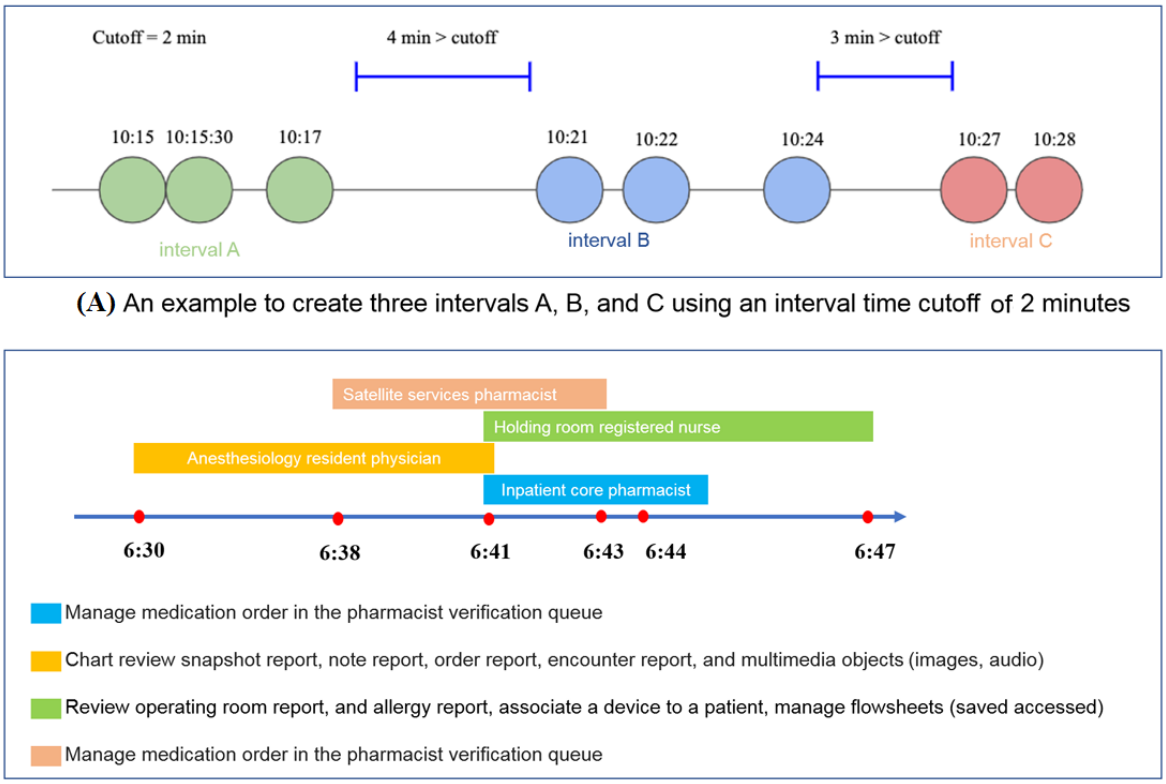
Table 1. Examples of events by health care workers.

Healthcare worker ID	Patient ID	Event action	Timestamp
A	1	FLOWSHEETS DATA SAVED	4/5/2020 2:14:25
A	1	CHART REVIEW ENCOUNTERS TAB SELECTED	4/5/2020 2:15:00
A	1	CHART REVIEW OTHER ORDERS TAB SELECTED	4/5/2020 2:18:23
A	1	HISTORY ACTIVITY ACCESSED	4/5/2020 2:19:53
A	1	FLOWSHEETS DATA COPIED FORWARD	4/5/2020 2:21:32
A	1	CHART REVIEW MEDICATIONS TAB SELECTED	4/5/2020 2:22:23
B	2	VISIT NAVIGATOR TEMPLATE LOADED	12/3/2020 06:31:27
B	2	SNAPSHOT REPORT VIEWED	12/3/2020 06:33:11
B	2	CHART REVIEW NOTES	12/3/2020 06:34:41
B	2	CHART REVIEW ENCOUNTER	12/3/2020 06:36:27
B	2	CHART REVIEW RESULTS	12/3/2020 06:37:33
B	2	CHART REVIEW OTHER ORDERS	12/3/2020 06:39:27

Creating Intervals From Events

We defined an interval as an ordered list of events that occur sequentially until two events are spaced in time by more than a certain cutoff (note that these events must be from the same HCW and the same patient). Each interval has start and stop times, corresponding to the first event and last event times in the interval. Intervals also have a duration metric, which is simply the difference between the start and stop times. Figure 2A provides a more detailed example using 2 minutes as a cutoff. This interval definition aims to divide an HCW's EHR actions into a set of segments, similar to an order session or a series of orders placed by a clinician for a single patient, defined in a previous report [33].

Figure 2. Examples of creating intervals from events (A) and defining a concurrent session based on overlapped intervals (B).


Cutoff = 2 min
4 min > cutoff
3 min > cutoff
10:15
10:15:30
10:17
10:21
10:22
10:24
10:27
10:28
interval A
interval B
interval C
Satellite services pharmacist
Holding room registered nurse
Anesthesiology resident physician
Inpatient core pharmacist
6:30
6:38
6:41
6:43
6:44
6:47
Manage medication order in the pharmacist verification queue
Chart review snapshot report, note report, order report, encounter report, and multimedia objects (images, audio)
Review operating room report, and allergy report, associate a device to a patient, manage flowsheets (saved accessed)
Manage medication order in the pharmacist verification queue

(A) An example to create three intervals A, B, and C using an interval time cutoff of 2 minutes

(B) An example of a concurrent session, consisting of four concurrent intervals

A “knee point” finding algorithm, described by Satopaa et al, was used to estimate the cutoff used [34]. Our previous study used such a strategy and identified clinically meaningful intervals for the sessionization of audit logs [35]. This strategy has been used before in finding the operating points of complex systems and is defined more formally, for any continuous function *f*, as follows:

$$K_f(x) = f''(x) / (1 + f'(x)^2)^{1.5} \quad \mathbf{(1)}$$

$K_f(x)$ represents the closed form of the curvature *f* at any point as a function of its first and second derivatives. We find *x* through the Kneedle algorithm, which maximizes this curvature [34].

Creating Concurrent Sessions From Intervals

We defined a concurrent session as a set of temporally overlapping intervals performed by different HCWs on EHRs of the same patient. We assumed that concurrent sessions can indicate who works with whom given that they are simultaneously performing EHR actions to manage a single patient. A concurrent interval is any interval that is part of a concurrent session; likewise, an individual interval is any interval that is not a part of any concurrent session. Concurrent intervals of a session have overlaps that are greater than zero. Figure 2B provides a more detailed example of a concurrent session made up of four concurrent intervals.

Workday Definition

We found that sometimes HCWs spend only a small amount of time (eg, 5 minutes per 24 hours) interacting with the EHRs of patients. We denoted these lower activity days as inactive EHR workdays and assumed that such workdays have little impact on measuring latent collaboration and its characteristics. Thus, we only investigated active EHR days in this study. An active EHR day was defined as a day (24 hours) where the sum of all the HCW interval durations in that day exceeds a certain amount of time, or the workday time cutoff. This cutoff value is determined by different clinical settings (eg, NICU or primary care) and the respective HCW time spent interacting with EHRs. We relied on expert knowledge in EHR utilization to determine the workday cutoff value.

Creating Intermediate Data Matrices

Based on the concurrent sessions, we generated eight intermediate matrices (Figure 1) describing latent collaboration and its characteristics. For instance, the associations between concurrent intervals and actions were stored in the concurrent interval-action matrix, and the associations between concurrent intervals and concurrent sessions were stored in the concurrent interval-concurrent session matrix. The intermediate data were used in the following analysis.

Measuring the Concurrent Intensity of an HCW

Given the definitions previously discussed, we can create attributes for each HCW. These attributes include HCW specialty (eg, neonatologist and neonatal nurse), individual intervals, concurrent intervals, durations of both individual and concurrent intervals, EHR workdays, and the durations of these EHR workdays. We leveraged these attributes to measure the proportion of concurrent intervals over all recorded intervals and the proportion of time spent on concurrent intervals. These two attributes comprise the concurrent intensity of an HCW. We measured the concurrent intensity per day, excluding inactive EHR workdays. The daily concurrent intensity, along with EHR time on active EHR workdays, was used to describe the time characteristics of an HCW in EHR systems. We used Spearman rank correlation to measure the association between daily time in EHRs and daily time spent on concurrent intervals for HCWs affiliated with the same specialty attribute. This tests the null hypothesis that there is no association between daily time spent on concurrent intervals and daily time spent on EHRs across all HCWs affiliated with the same specialty. Moreover, we applied a one-way analysis of variance (ANOVA) to test the significance of differences in the concurrent intensity and EHR time on active workdays between specialties at a significance level of .05. The null hypothesis is that there are no significant differences in the concurrent intensity/EHR time between HCWs with disparate specialties. All statistical analyses, including those in the following sections, were performed using R 4.0.4 (R Foundation for Statistical Computing). The four matrices, as shown in Figure 1, were used to quantify concurrent intensity for each HCW.

Measuring and Validating Latent Collaborative Relationships Between HCWs

The HCW-concurrent session matrix was leveraged to measure relationships between HCWs with respect to their participation in concurrent sessions. In this study, we used the number of co-affiliated sessions between pairs of HCWs to measure the relationship's strength. Based on these weightings between HCWs, we created a network of HCWs to describe their latent collaborations. We used K-core analyses to identify a subgraph depicting core latent collaboration among HCWs in EHR systems. Each HCW within the K-core subgraph is connected to at least K other HCWs, and each respective HCW is considered as one core of the whole collaboration network. Gephi, an open-sourced network analysis and visualization tool, was used in this study [36].

We assumed that if the learned classes of the collaborative relationships (high and low strength) are consistent with the psychological expectations of HCWs, our approaches measuring latent collaborative relationships are plausible. To assess if HCWs can distinguish between likelihoods of collaborative relationships derived from EHRs, we divided putative collaborative relationships into the following two groups: high and low likelihoods. We randomly selected a set of collaborative relationships from the high and low groups, which were assessed by invited experts in an online survey. The experts who responded to the survey were asked questions like "To what extent do you believe [a neonatal nurse] interacts with [a neonatologist] in the electronic health record system to manage a patient?" This is asked for each collaborative relationship, and respondents are blind to the EHR-learned likelihood. The professionals were asked to choose one of the following five answers: "Not at all likely," "Slightly likely," "Moderately likely," "Very likely," and "Completely likely." For statistical analysis, these survey responses were encoded as integer values (Likert score) in the range 1 to 5 (eg, "Not at all likely" is mapped to 1). The Likert scores were used to quantify an expert's psychological expectations of latent collaborative relationships.

These surveys were distributed through the REDCap management system [37] and expert responses were requested after review and approval from the Vanderbilt Institutional Review Board (approval number: 191892). Using the survey results, we tested the following hypothesis: experts can distinguish latent collaborative relationships between high and low likelihood categories. We applied a linear regression model, shown in the following equation, to determine the Likert score for high and low likelihood relationships.

$$\text{Likert Score}=\alpha+\theta\times\beta \quad \textbf{(2)}$$

where θ {1 (high likelihood), 0 (low likelihood)} represents the high and low likelihoods of collaborative relationships identified from EHRs. Under this model, the Likert score for a low likelihood collaboration is α ($\theta=0$) and for a high likelihood collaboration is $\alpha + \beta$ ($\theta=1$). As such, the value of β corresponds to the difference of Likert scores for high and low likelihood collaborative relationships.

We used the Likert scores as observations to infer β via linear regression models. We then used ANOVA to test the significance of $\beta\neq0$ against a null hypothesis $\beta=0$. We tested the hypothesis at the two-sided $\alpha=.05$ significance level.

Mining Weekday and Weekend Temporal Trends of Concurrent EHR Usage

We analyzed when (eg, shifts) concurrent sessions occur in EHRs. We assumed HCWs have different EHR interaction patterns during weekdays and weekends, and that those patterns are also different in the phases of a patient's stay in the NICU. Therefore, we modeled weekday- and weekend-temporal trends of concurrent EHR usage, which we called temporal patterns, and focused on the following three specific phases of a patient's hospital stay: admission, discharge, and intermediate phases.

Since all investigated patients had admission and discharge dates, we learned temporal patterns 24 hours after admission and 24 hours before discharge based on all those patients. We created the following four patient groups: (1) patients admitted during weekdays, (2) patients admitted during weekends, (3) patients discharged during weekdays, and (4) patients discharged during weekends. For a single patient in a patient group, we measured the number of concurrent intervals performed by HCWs on EHRs of that patient in each hour during the 24-hour window. Next, we calculated the average number in each hour for all patients in a group to form a temporal pattern. We used Wilcoxon rank-sum test to measure differences in the temporal patterns between weekdays and weekends because the Wilcoxon rank-sum test is based solely on the order in which the observations from the two patterns fall.

We chose days surrounding the middle of a patient's hospital stay to represent the intermediate phase for the measurement of temporal patterns. We also separated patient stay into the following two subgroups: weekday and weekend stay. We measured the average number of concurrent intervals for each hour and used the Wilcoxon rank-sum test to assess the differences between weekday and weekend temporal patterns.

We also compared the differences between weekdays and weekends in the degree of concurrent EHR usage during the admission, discharge, and intermediate phases of hospital stay using *t* tests. This was done to determine if there was a significant difference between the means of two patterns, without the consideration of pattern observation order.

Clustering Concurrent Intervals to Describe a Concurrent Session

The concurrent interval-action matrix recorded the number of times an action appeared in a concurrent interval. This matrix was used to learn similarities between concurrent intervals and concurrent sessions in terms of their affiliated action types. We performed principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and K-means clustering to aggregate intervals described by a cluster-concurrent interval matrix. The cluster-concurrent interval matrix was used jointly with the concurrent interval-concurrent session matrix to determine if a concurrent session contains intervals assigned to the same cluster or different clusters. This joint analysis was performed by calculating the dot product of the matrices. Such an analysis can highlight the complexity (eg, a concurrent session affiliated with a single cluster or multiple clusters) of a concurrent session.

Availability of Data

The data sets that were generated and analyzed in this study are not publicly available because they include patients' private information. However, the data sets can be obtained from the corresponding author upon reasonable request.

Results

Case Studies in the NICU

We gathered 4 months of EHR audit log data from a large academic medical center's NICU. The data set contained 2,840,249 actions performed by 3303 HCWs (approximately 22,319 HCWs in the VUMC EHR system) to EHRs of 382 NICU patients. In this case study, we identified 2 minutes as the cutoff threshold in creating intervals from series of events; this was the point of maximum curvature, or “knee point,” determined through the Kneedle algorithm (as shown in Multimedia Appendix 1). We used 15 minutes as the threshold to separate inactive and active workdays, as determined through NICU expert questionnaires regarding EHR utilization. From these thresholds, we created 624,192 concurrent intervals, each of which comprised of consecutive sequences of 650 unique actions. There were 173,436 concurrent sessions created from the concurrent intervals.

Examining Differences in the Concurrent Intensity Between NICU HCWs

We compared the concurrent intensities across the top 13 specialties having the highest average EHR times on active workdays. The 13 specialties consisting of 552 HCWs are listed in Table 2, along with their mean values and 95% CIs for concurrent intensity and EHR time. The concurrent intensity was calculated after excluding activities on inactive workdays. The statistical test results showed that there were significant differences in the EHR time (from mean 23.38, 95% CI 21.97-24.80, to mean 54.78, 95% CI 40.43-69.13; *P*<.001) and concurrent intensity (from mean 0.07, 95% CI 0.06-0.08, to mean 0.36, 95% CI 0.18-0.54; *P*<.001 with respect to the proportion of concurrent intervals and from mean 0.32, 95% CI 0.20-0.44, to mean 0.76, 95% CI 0.51-1.00; *P*<.001 with respect to the proportion of EHR time spent on concurrent intervals) between the 13 investigated specialties. We found that there were no significant relationships between EHR time and the proportion of time spent on concurrent intervals, except for extracorporeal membrane oxygenation (ECMO) respiratory therapists (*P*<.001). ECMO respiratory therapists had positive associations between time spent in EHRs and the proportion of time spent on concurrent intervals. This indicates that ECMO respiratory therapists work (76% of their EHR time) in a highly concurrent environment.

Table 2. Data for health care workers affiliated with 13 specialties.

Specialty	EHR[a] time (min), mean (95% CI)	Number of event actions per day, mean (95% CI)	Proportion of concurrent intervals, mean (95% CI)	Proportion of EHR time spent on concurrent intervals, mean (95% CI)
MRI[b]-technologists	54.78 (40.43-69.13)	66.49 (54.82-78.16)	0.25 (0.07-0.44)	0.51 (0.17-0.85)
Diagnostic radiology-technologists	50.06 (39.77-60.35)	72.53 (62.64-82.41)	0.36 (0.18-0.54)	0.60 (0.39-0.80)
Pediatric cardiac ICU[c]-registered nurse	49.35 (45.02-53.69)	144.03 (134.77-153.29)	0.13 (0.08-0.18)	0.59 (0.48-0.69)
NICU[d]-registered nurse	36.56 (35.17-37.95)	98.86 (96.95-100.77)	0.07 (0.06-0.08)	0.40 (0.37-0.44)
Pediatrics-resident physician	34.03 (30.71-37.36)	121.58 (114.72-128.45)	0.08 (0.05-0.11)	0.63 (0.55-0.71)
Float pool-registered nurse	33.47 (26.06-40.88)	86.33 (75.05-97.62)	0.09 (0.02-0.15)	0.32 (0.20-0.44)
Inpatient-nurse practitioner	31.36 (28.75-33.98)	131.74 (124.82-138.66)	0.14 (0.09-0.18)	0.66 (0.58-0.75)
ECMO[e]-registered nurse	30.73 (21.93-39.52)	91.88 (75.31-108.46)	0.21 (0.04-0.38)	0.70 (0.39-1.00)
ECMO-respiratory therapist	30.62 (22.58-38.65)	93.31 (77.85-108.77)	0.28 (0.06-0.50)	0.76 (0.51-1.00)
Perioperative services-registered nurse	27.89 (20.50-35.27)	90.43 (58.77-122.08)	0.16 (0.07-0.26)	0.68 (0.40-0.96)
Rx inpatient core-pharmacist	26.27 (20.22-32.33)	58.16 (53.18-63.14)	0.12 (0.07-0.18)	0.75 (0.52-0.98)
Anesthesiology-nurse anesthetist	24.91 (19.86-29.96)	105.08 (88.39-121.78)	0.15 (0.06-0.23)	0.63 (0.38-0.89)
Pediatrics-respiratory therapist	23.38 (21.97-24.80)	103.87 (98.58-109.16)	0.11 (0.07-0.15)	0.57 (0.46-0.69)

[a] EHR: electronic health record.

[b] MRI: magnetic resonance imaging.

[c] ICU: intensive care unit.

[d] NICU: neonatal intensive care unit.

[e] ECMO: extracorporeal membrane oxygenation.

Examining Latent Collaboration Networks in the NICU

We identified a collaboration network consisting of 857 HCWs with 4242 edges connecting them. The 857 HCWs were affiliated with 406 unique specialties. Figures 3 and 4 show the collaboration network of HCWs and its 15-core subnetwork, where each node is an HCW. The 15-core subnetwork was made up of 61 core HCWs, with 748 edges connecting those HCWs. Within the 15-core subnetwork, each HCW collaborated with at least 15 other HCWs. Compared with the full collaborative network (centered by a neonatal nurse), the 15-core subnetwork was centered by ancillary staff (latent collaboration with NICU professionals: neonatal nurses, neonatal frontline providers, neonatal fellows, neonatologists, and respiratory therapists). To interpret these collaborations among NICU HCWs, NICU experts categorized 406 specialties into the following eight roles: neonatologists, neonatal fellows, neonatal frontline providers (eg, nurse practitioners, physician assistants, hospitalists, and resident physicians), neonatal nurses, respiratory therapists, consultants (eg, surgeons, OB/GYN physicians, hematology physicians, radiology physicians, anesthesiologists, and genetics counselors), ancillary staff (eg, registered dietitians, social workers, case managers, technicians, and phlebotomists), and support staff (eg, clerks, information technology staff, coordinators, and medical assistants). The collaboration network, as shown in Figure 5, was visualized at the level of roles. The mappings between the eight roles and 406 specialties are presented in Multimedia Appendix 2.

Overall, ancillary staff and consultants had higher concurrent intensity (larger node size) than HCWs with other roles (Figure 3). Supporting this was the observation that ancillary staff nodes were distributed across the 15-core subnetwork (Figure 4). Figure 5 indicates the latent collaborative relationships among the eight professional roles. The relationships between ancillary HCWs, consultants, neonatal nurses, and neonatal frontline providers were strong. Neonatal nurses were very active in the network, often collaborating with HCWs from ancillary staff, which was the most collaborative role.

Figure 3. The latent collaboration network of HCWs. Each HCW is coded as a color based on their affiliated role category. The size of the node is determined by the proportion of time spent on the concurrent intervals over all intervals. A larger node size is associated with a higher proportion of time for concurrent EHR usage. EHR: electronic health record; HCW: health care worker.

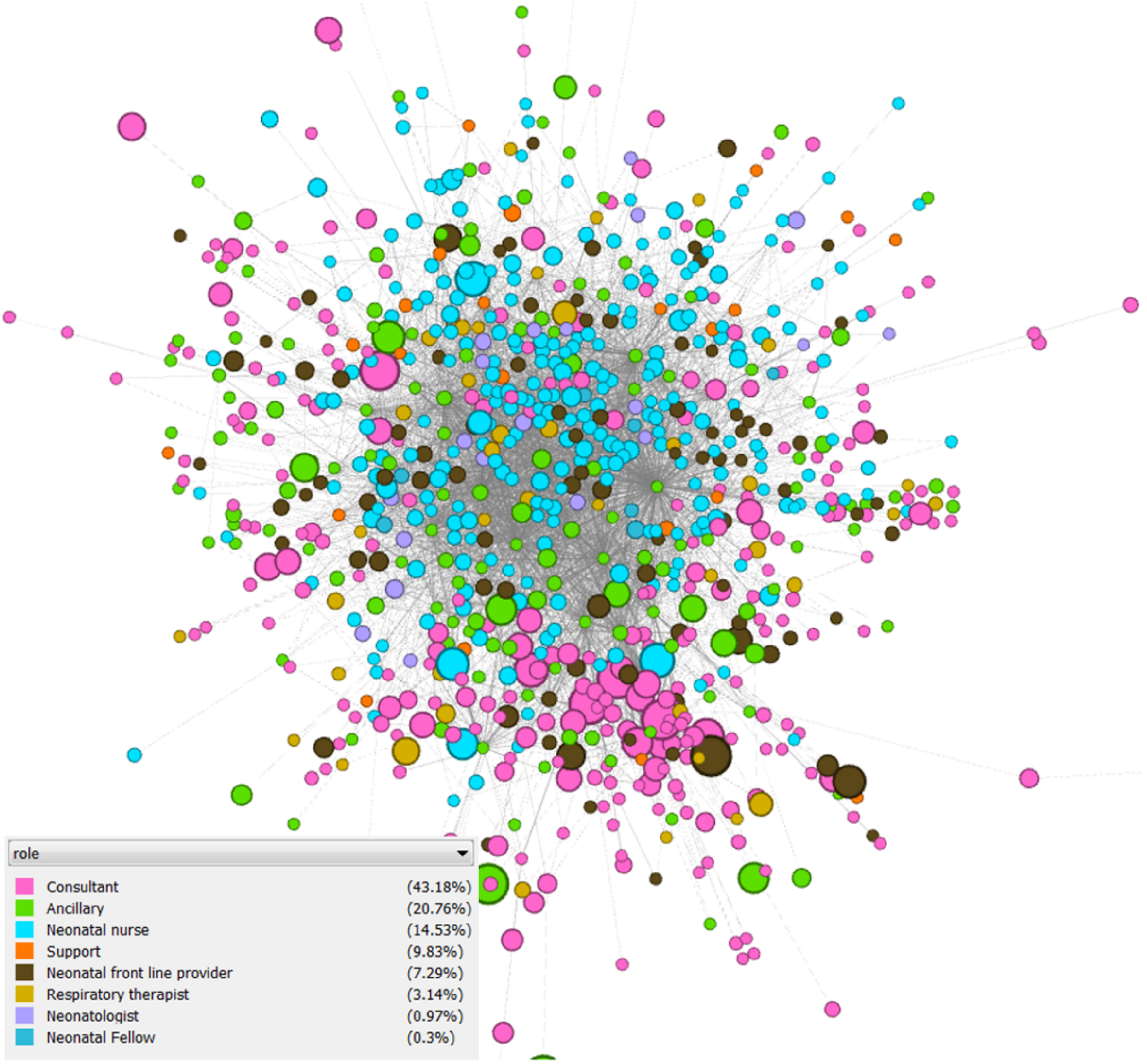


Figure 4. The 15-core HCW subnetwork. Each node is an HCW labeled by the roles. The size of the node is determined by the proportion of time spent on the concurrent intervals over all intervals. A larger node size is associated with a higher proportion of time for concurrent EHR usage. EHR: electronic health record; HCW: health care worker.

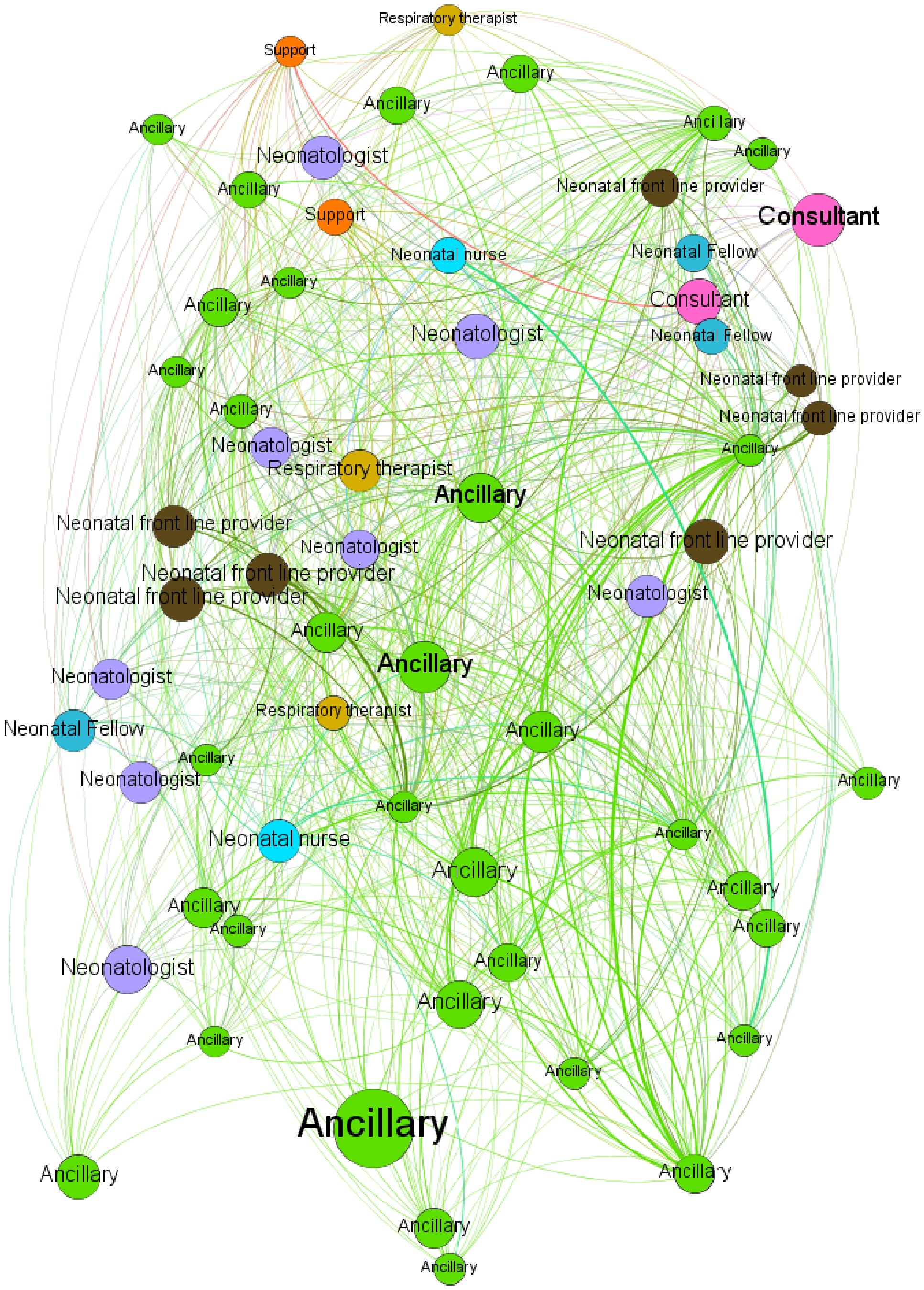
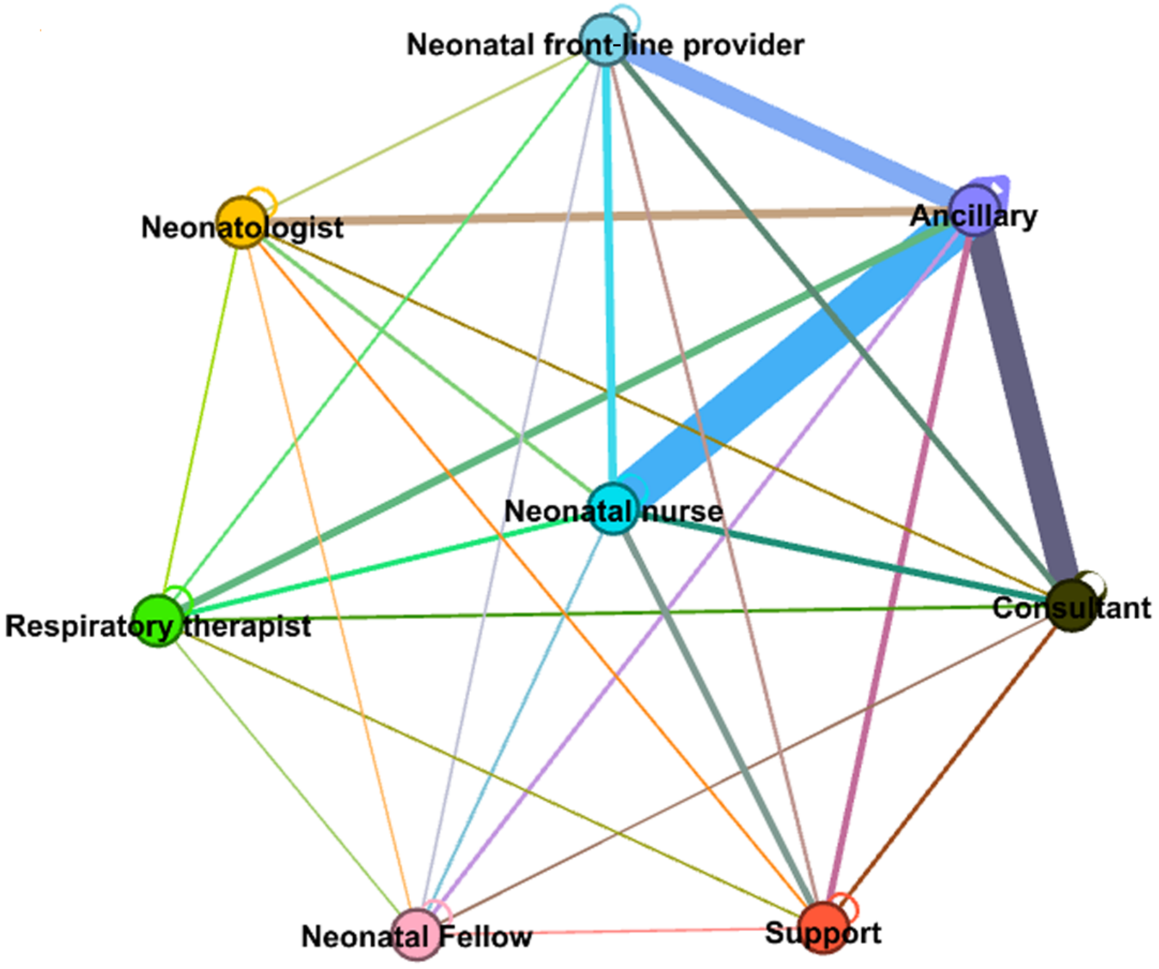


Figure 5. The collaboration network of the eight role categories. The nodes are roles. The weight of the edge indicates the strength of the collaboration.



Collaboration Validation Results

We sampled 12 (12/28) collaborative relationships (six of high and six of low likelihood), and generated a survey containing 12 questions (Multimedia Appendix 3). The total number of NICU experts who accepted the invitation and participated in the survey was 13, with four neonatologists, three neonatal fellows, three neonatal nurses, two nurse practitioners, and one respiratory therapist. The 13 experts, including neonatal attendings, nurses, nurse practitioners, residents, fellows, and respiratory therapists, were representatives of the expertise in the NICU. The average number of years those experts have been working at the NICU is 5.65. All 13 responding NICU experts completed the survey (100% response rate). The number of years of experts working in the NICU is depicted in Multimedia Appendix 4. The results of the Likert scores are shown in Multimedia Appendix 5. Our assumption of using a linear regression model was confirmed by the quantile-quantile plot, as shown in Figure S2 in Multimedia Appendix 5. Overall, NICU experts could distinguish collaborative relationships between high and low likelihoods (β=.88, Likert score: 3.54, 95% CI 3.31-4.37 vs 2.64, 95% CI 2.46-3.29; *P*<.001). The Likert scores of the 12 collaborations surveyed from NICU experts are shown in Table 3.

Table 3. Likert scores of the 12 investigated collaborative relationships.

Index and collaborative relationship	Likert score from NICU[a] professionals
Collaborative relationships with high likelihoods learned from EHRs[b]	
1: Neonatal front line provider ↔ Consultant	3.85
2: Ancillary staff ↔ Consultant	2.23
3: Neonatal nurse ↔ Consultant	4.46
4: Neonatal front line provider ↔ Ancillary staff	4.08
5: Ancillary staff ↔ Neonatal nurse	3.46
6: Support staff ↔ Neonatal nurse	3.15
Collaborative relationships with low likelihoods learned from EHRs	
7: Neonatal front line provider ↔ Neonatologist	3.15
8: Ancillary staff ↔ Support staff	2.07
9: Neonatal nurse ↔ Neonatal fellow	3.07
10: Neonatal front line provider ↔ Neonatal fellow	2.84
11: Ancillary staff ↔ Neonatal fellow	3.07
12: Support staff ↔ Neonatal fellow	1.69

[a] NICU: neonatal intensive care unit.

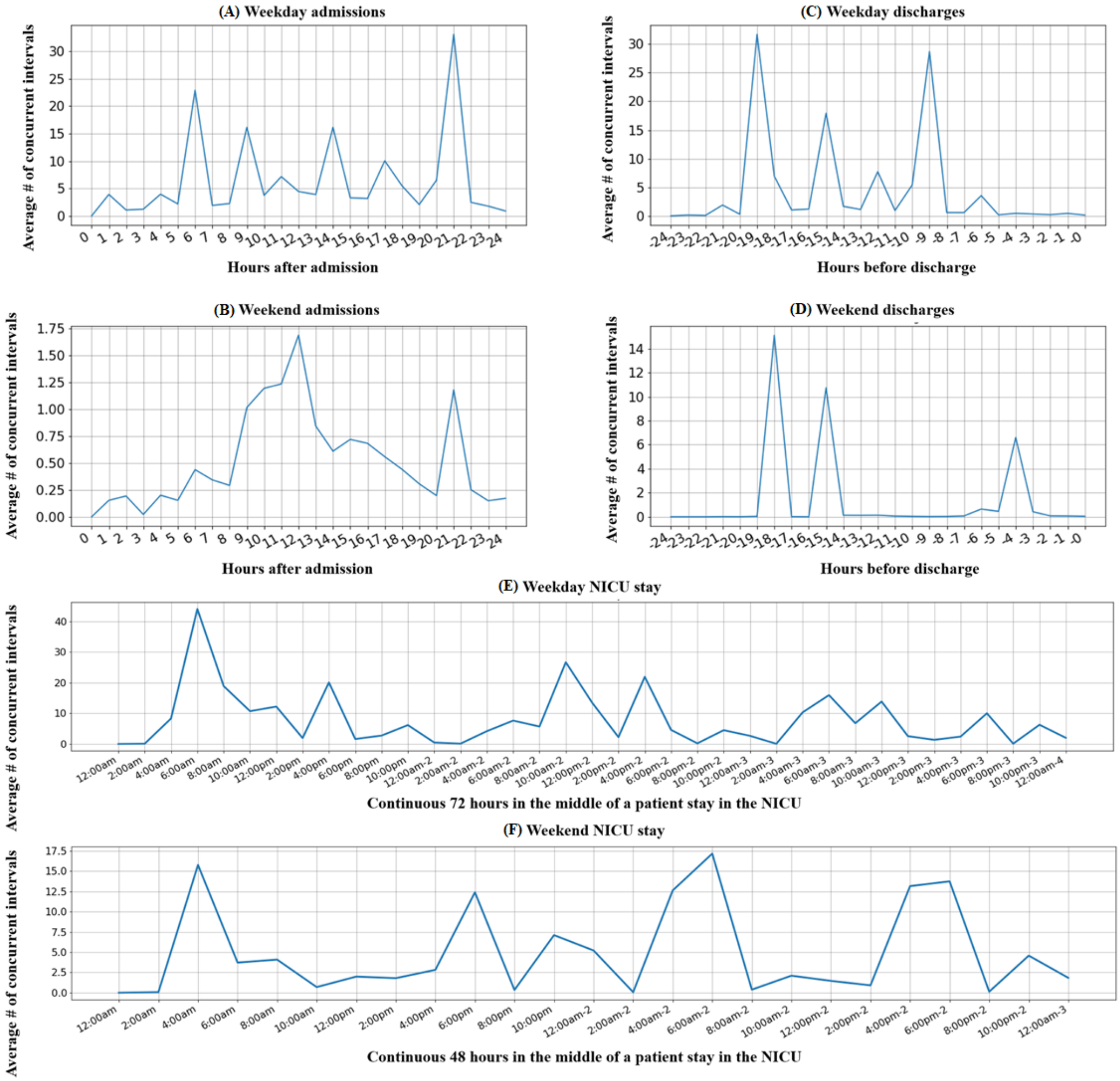
[b] EHR: electronic health record.

Temporal Trends of Concurrent EHR Usage

Figure 6 shows the temporal patterns 24 hours after admission (6A and 6B) and 24 hours before discharge (6C and 6D). These patterns were separated by weekday or weekend status. The temporal patterns were significantly different for both admission (average number of concurrent intervals per hour: 11.60 vs 0.54, *P*<.001) and discharge days (4.72 vs 1.45, *P*<.001), but not for the intermediate phase of hospital stay.

Expectedly, there was more concurrent EHR usage between HCWs on weekdays than weekends across all three phases of hospital stay (average number of concurrent intervals per hour: 9.56 vs 2.34, *P*<.001).

Figure 6. Concurrent EHR usage temporal trends 24 hours after admission (A and B), 24 hours before discharge (C and D), and consecutive intermediate days of hospital stay. The trends are measured from weekdays and weekends. EHR: electronic health record; NICU: neonatal intensive care unit.



Clusters of Concurrent Intervals and the Composition of a Concurrent Session

We clustered the concurrent intervals using their constituent actions as features. We used PCA to reduce the dimensionality to the top 10 components, which explained 97% of the variance. We then applied t-SNE on the 10 PCA components to further reduce the data to two dimensions. Finally, we used the k-means clustering algorithm to form 50 clusters, as shown in Figure 7. This K of 50 was determined by minimizing the total within-cluster sum of squared errors (WSS). The squared error for each point was the square of the distance of the point from its predicted cluster center. The WSS score was the sum of these squared errors for all the points. The plot of WSS versus k is depicted in Figure S1 in Multimedia Appendix 6. As shown in Figure 7, the 50 clusters were well separated. Concurrent intervals within each cluster shared similar actions.

Using the clusters visualized in Figure 7, we examined intercluster relationships, as shown in Figure S2 in Multimedia Appendix 6. The calculated intercluster network described the pairwise relationship of each of the concurrent sessions.

Figure 8 shows the distribution of concurrent sessions in terms of the number of clusters affiliated. We showed that over 87% of concurrent sessions could be unambiguously assigned into a single unique cluster, indicating that most HCWs perform similar actions in a concurrent session. About 13% of concurrent sessions, consisting of concurrent intervals, came from multiple clusters.

Our unsupervised learning framework could identify and quantify concurrent EHR usage from audit log data. Based on concurrent EHR usage, we could determine the proportion of concurrent activities, the proportion of time spent on those activities, HCWs who participate in concurrent or latent interactions, the temporal trends of concurrent EHR usage on weekdays and weekends in the three phases of hospital stay, and the complexity of concurrent activities (single cluster vs multiple clusters).

Figure 7. A visualization of the 50 clusters of concurrent intervals. Each node is a concurrent interval, and each color indicates the cluster group to which an interval belongs. The axes are t-distributed stochastic neighbor embedding–reduced components.

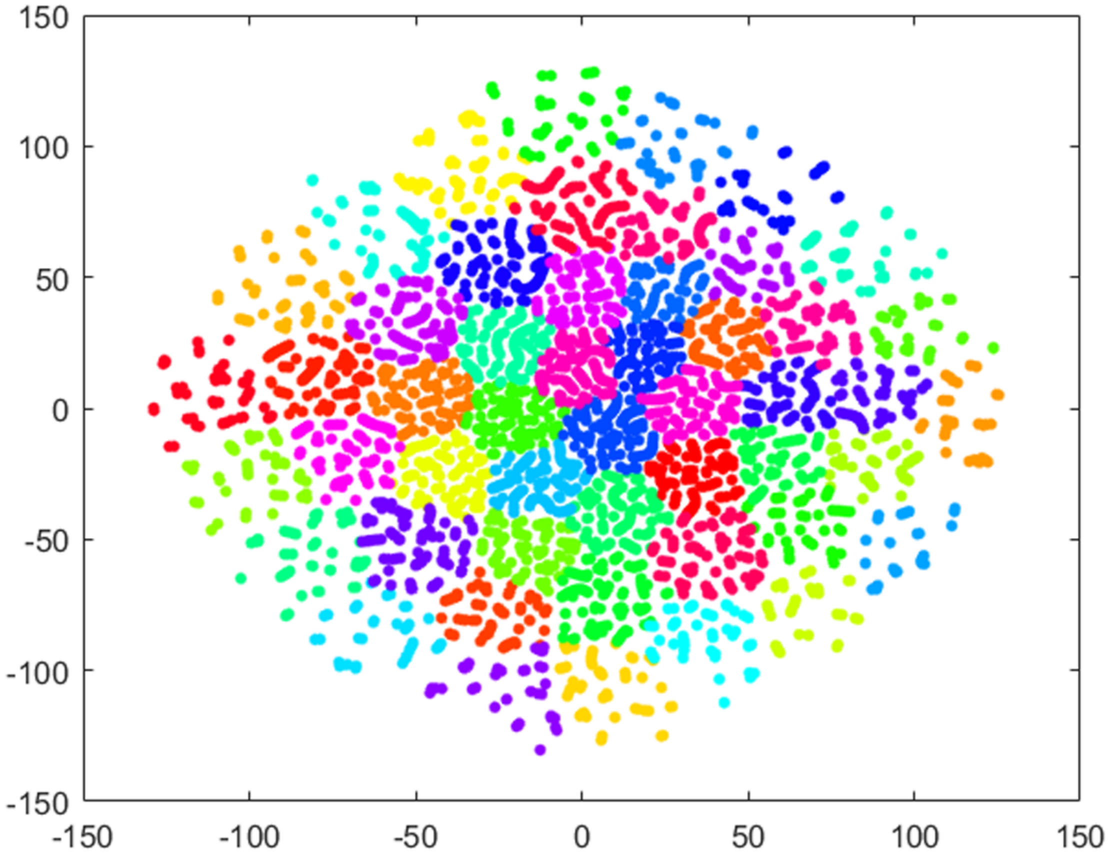
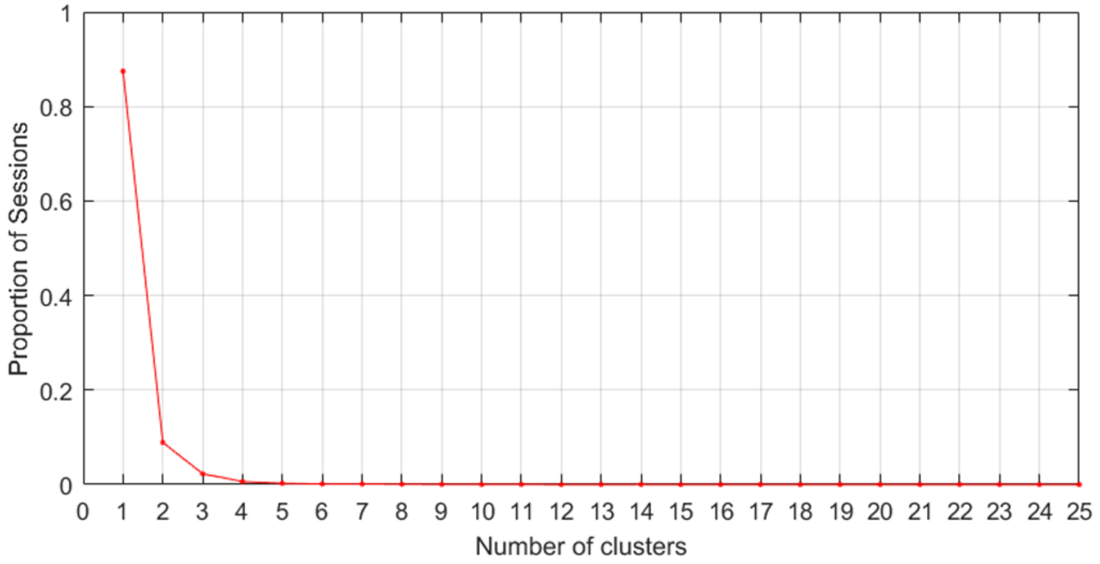


Figure 8. The distribution of concurrent sessions as a function of the number of clusters that concurrent intervals are affiliated with.



Discussion

Principal Findings

We presented a novel framework to measure latent collaboration from EHR audit logs, and we established novel metrics, which may be useful for the analysis of latent HCW collaboration. EHR system usage is pervasive and still increasing. While there are studies that measured collaboration, few targeted the growing paradigm of latent collaboration among HCWs. We demonstrated the use of our informatics framework in the analysis of latent collaboration. We examined the concurrent intensity across various HCW specialties and found that there was a statistically significant difference in the proportion of concurrent activities and the proportion of time spent on those activities. It was noted that in some settings, clinicians shared the same workstation or computer terminal. Concurrent EHR usage may have highly variable ergonomics between health care settings, for example, in some instances, HCWs may have to share one workstation, making concurrent EHR usage impossible. In this study, we identified latent collaboration among HCWs coming from various departments, and thus, there was a low probability for HCWs sharing the same workstation or computer terminal. If latent collaboration is identified among HCWs from the same department or unit, it would be better for HCOs to allocate more workstations or computer terminals to HCWs within the department/unit to achieve high performing collaboration in EHR systems.

We examined networks that represented the collaborative relationships between HCWs (Figure 2). By using our framework, we identified HCW relationships between defined role categories in the NICU. We assessed our framework in a NICU setting, and it demonstrated the effectiveness of using concurrent EHR usage measuring latent collaboration. Based on the observations from Figures 3 to 5, EHR vendors or HCOs may need to establish communication channels in EHR systems for ancillary staff to collaborate with other HCWs (eg, NICU nurses) to deliver high quality care for neonates.

Strikingly, strong collaborative relationships between consultants, ancillary staff, and neonatal nurses are described by our framework (Figure 4), though NICU experts do not consistently assign collaborative relationships between them (eg, collaborative relationship between ancillary staff and consultants) (Table 3). One potential reason for this discrepancy is that our survey respondents were not part of ancillary staff or consultant roles, thus limiting the description of these specific collaborations. Recruiting HCWs from these roles as survey respondents remains high priority, but is challenging due to their assignments to heterogeneous departments and care units. We believe a large scale study is required to formally assess latent collaborative relationships between ancillary staff and consultants.

We examined concurrent EHR usage patterns in the admission, discharge, and intermediate phases of hospital stay, finding significant differences in patterns between weekdays and weekends. This suggests that HCWs act differently on weekdays and weekends, which may assist HCOs in using different staffing strategies optimizing latent collaboration on weekdays and weekends.

We clustered the concurrent intervals of HCWs and highlighted their interconnectivity (Multimedia Appendix 6). These clusters and their neighbors may be used to reduce the search space for the analysis of audit log data. Potentially, this enables higher throughput process mining or the targeting of specific dominant HCW roles.

Scope of This Study and Its Limitations

This was a pilot study, and we would like to acknowledge some limitations that may guide prospective latent collaboration-related studies. Using concurrent HCW activity can help HCOs or EHR vendors identify potential collaborative relationships among HCWs; however, such relationships need to be further validated when optimizing or refining EHR systems. Moreover, causative explanations for these latent relationships are not determined. We believe that describing the causes for certain collaborations would require additional data and further investigations on the HCW-EHR system interaction workflow. This study does not describe the cause of this phenomenon, but highlights its existence and provides an avenue of hypothesis generation for future work.

There are multiple forms of collaboration between HCWs [26]. Collaboration may consist of direct and explicit physical communication or latent interactions through digital platforms, but our study focused on latent interactions involving EHR systems. Learning broader forms of collaboration requires the integration of a broader range of data resources.

We investigated when concurrent EHR usage occurs, but did not investigate the underlying causes for the observed differences. Our focus on concurrent EHR usage may not be able to detect collaborative activities that do not have time overlaps. Further, we acknowledge that not every piece of concurrent HCW activity indicates a latent collaboration. It is possible that overlapping usage of the same target patient EHR is coincidental. For instance, some HCWs may simply have overlapping shifts, which may be detected as false positives with our framework, thus requiring further validation to flag these scenarios.

Since interval durations were calculated through the difference of timestamps, we did not capture the duration of interval-ending actions. Potential remedies in logging the durations of these types of actions include the use of video monitoring to track HCW activities in EHRs.

NICU experts distinguished latent collaborative relationships between high and low likelihoods learned from EHRs; however, we did not assess the plausibility of each inferred latent collaborative relationship at the level of the EHR user (edges in Figure 3).

Finally, we used a threshold determined by experts to define active EHR workdays. Activities occurring on inactive EHR workdays may also contribute evidence for measuring latent collaborative relationships. Moreover, categorization of 406 specialists named by the Epic system into eight general roles was conducted by NICU experts, which may be biased according to their expertise and experiences.

Conclusion

We presented an informatics framework relying on concurrent EHR usage to learn latent collaboration. We explored the advantages of the framework by conducting the following four types of analyses: (1) quantifying time spent interacting with EHRs and on concurrent usage, (2) investigating the latent collaborative relationships among HCWs engaging in highly concurrent EHR usage, (3) measuring temporal trends of concurrent EHR usage on weekdays and weekends in the three phases of hospital stay, and (4) clustering EHR activities to describe the complexity of concurrent EHR usage. We assessed the effectiveness of our framework through a case study and anticipated that its generalizability will further enable the analysis of how latent collaborative interactions affect patient care, discharge times, and clinician workload, stress, or burnout.

Acknowledgments

This research was supported, in part, by the National Library of Medicine of the National Institutes of Health under award numbers R01LM012854 and T32LM012412, and the National Science Foundation under award number 1757644. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and National Science Foundation.

Authors' Contributions

PL and YC conceived the presented idea and performed data collection and analysis, method and metric design and development, experiment design, evaluation and interpretation of the experiments, and writing and revision of the manuscript. ER and MWA invited experts to complete online surveys, performed evaluation and interpretation of the experiments, and revised the manuscript. BC, JS, and DF performed evaluation and interpretation of the experiments, and revised the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Determination of the interval time threshold.
[DOCX File , 140 KB - medinform_v9i9e28998_app1.docx]

Multimedia Appendix 2
Categorization of specialists into roles.
[DOCX File , 40 KB - medinform_v9i9e28998_app2.docx]

Multimedia Appendix 3
Survey questions.
[DOCX File , 24 KB - medinform_v9i9e28998_app3.docx]

Multimedia Appendix 4
Summary statistics of neonatal intensive care unit experts participating in online surveys.
[DOCX File , 21 KB - medinform_v9i9e28998_app4.docx]

Multimedia Appendix 5
Likert scores of 12 collaboration relationships surveyed from neonatal intensive care unit experts, and the plot of standardized residuals for the linear regression model.
[DOCX File , 34 KB - medinform_v9i9e28998_app5.docx]

Multimedia Appendix 6
Determination of the number of clusters, and visualization of the cluster-cluster relationships.
[DOCX File , 142 KB - medinform_v9i9e28998_app6.docx]

References

1. Costa DK, Valley TS, Miller MA, Manojlovich M, Watson SR, McLellan P, et al. ICU team composition and its association with ABCDE implementation in a quality collaborative. J Crit Care 2018 Apr;44:1-6 [FREE Full text] [doi: 10.1016/j.jcrc.2017.09.180] [Medline: 28978488]
2. Lilly CM, Mullen M. Critical Care Surge Management. Critical Care Medicine 2019;47(9):1271-1273. [doi: 10.1097/ccm.0000000000003881]
3. Donovan AL, Aldrich JM, Gross AK, Barchas DM, Thornton KC, Schell-Chaple HM, et al. Interprofessional Care and Teamwork in the ICU. Critical Care Medicine 2018;46(6):980-990. [doi: 10.1097/ccm.0000000000003067]
4. Bunkenborg G, Bitsch Hansen T, Hølge-Hazelton B. Handing over patients from the ICU to the general ward: A focused ethnographical study of nurses' communication practice. J Adv Nurs 2017 Dec 14;73(12):3090-3101. [doi: 10.1111/jan.13377] [Medline: 28677173]

5. Despins LA. Patient safety and collaboration of the intensive care unit team. Crit Care Nurse 2009 Apr;29(2):85-91. [doi: 10.4037/ccn2009281] [Medline: 19339450]
6. Pronovost PJ, Berenholtz SM, Goeschel C, Thom I, Watson SR, Holzmueller CG, et al. Improving patient safety in intensive care units in Michigan. J Crit Care 2008 Jun;23(2):207-221. [doi: 10.1016/j.jcrc.2007.09.002] [Medline: 18538214]
7. Rose L. Interprofessional collaboration in the ICU: how to define? Nurs Crit Care 2011;16(1):5-10. [doi: 10.1111/j.1478-5153.2010.00398.x] [Medline: 21199549]
8. Nap R, Silva Alvaro M, Fidler V, Reis Miranda D. Collaborative practice and clinical outcomes in the ICU. Critical Care 2000;4(Suppl 1):P221. [doi: 10.1186/cc940]
9. Chunchu K, Mauksch L, Charles C, Ross V, Pauwels J. A patient centered care plan in the EHR: improving collaboration and engagement. Fam Syst Health 2012 Sep;30(3):199-209. [doi: 10.1037/a0029100] [Medline: 22866953]
10. Doupi P. Using EHR data for monitoring and promoting patient safety: reviewing the evidence on trigger tools. Stud Health Technol Inform 2012;180:786-790. [Medline: 22874299]
11. Kutney-Lee A, Kelly D. The effect of hospital electronic health record adoption on nurse-assessed quality of care and patient safety. J Nurs Adm 2011 Nov;41(11):466-472 [FREE Full text] [doi: 10.1097/NNA.0b013e3182346e4b] [Medline: 22033316]
12. Begum R, Smith Ryan M, Winther C, Wang J, Bardach N, Parsons A, et al. Small practices' experience with EHR, quality measurement, and incentives. Am J Manag Care 2013 Nov;19(10 Spec No):eSP12-eSP18 [FREE Full text] [Medline: 24511883]
13. Durojaiye AB, Levin S, Toerper M, Kharrazi H, Lehmann H, Gurses A. Evaluation of multidisciplinary collaboration in pediatric trauma care using EHR data. J Am Med Inform Assoc 2019 Jun 01;26(6):506-515 [FREE Full text] [doi: 10.1093/jamia/ocy184] [Medline: 30889243]
14. Chase DA, Ash J, Cohen D, Hall J, Olson G, Dorr D. The EHR's roles in collaboration between providers: A qualitative study. AMIA Annu Symp Proc 2014;2014:1718-1727 [FREE Full text] [Medline: 25954444]
15. Reitz R, Common K, Fifield P, Stiasny E. Collaboration in the presence of an electronic health record. Fam Syst Health 2012 Mar;30(1):72-80. [doi: 10.1037/a0027016] [Medline: 22429079]
16. Kim C, Lehmann C, Hatch D, Schildcrout J, France D, Chen Y. Provider Networks in the Neonatal Intensive Care Unit Associate with Length of Stay. IEEE Conf Collab Internet Comput 2019 Dec;2019:127-134 [FREE Full text] [doi: 10.1109/CIC48465.2019.00024] [Medline: 32637942]
17. Overhage JM, McCallie D. Physician Time Spent Using the Electronic Health Record During Outpatient Encounters. Ann Intern Med 2020 Jan 14;172(3):169. [doi: 10.7326/m18-3684]
18. Goldstein IH, Hwang T, Gowrisankaran S, Bales R, Chiang MF, Hribar MR. Changes in Electronic Health Record Use Time and Documentation over the Course of a Decade. Ophthalmology 2019 Jun;126(6):783-791 [FREE Full text] [doi: 10.1016/j.ophtha.2019.01.011] [Medline: 30664893]
19. Varpio L, Rashotte J, Day K, King J, Kuziemsky C, Parush A. The EHR and building the patient's story: A qualitative investigation of how EHR use obstructs a vital clinical activity. Int J Med Inform 2015 Dec;84(12):1019-1028. [doi: 10.1016/j.ijmedinf.2015.09.004] [Medline: 26432683]
20. Rule A, Chiang M, Hribar M. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. J Am Med Inform Assoc 2020 Mar 01;27(3):480-490 [FREE Full text] [doi: 10.1093/jamia/ocz196] [Medline: 31750912]
21. Adler-Milstein J, Adelman JS, Tai-Seale M, Patel VL, Dymek C. EHR audit logs: A new goldmine for health services research? J Biomed Inform 2020 Jan;101:103343 [FREE Full text] [doi: 10.1016/j.jbi.2019.103343] [Medline: 31821887]
22. Amroze A, Field TS, Fouayzi H, Sundaresan D, Burns L, Garber L, et al. Use of Electronic Health Record Access and Audit Logs to Identify Physician Actions Following Noninterruptive Alert Opening: Descriptive Study. JMIR Med Inform 2019 Mar 07;7(1):e12650 [FREE Full text] [doi: 10.2196/12650] [Medline: 30730293]
23. Wang JK, Ouyang D, Hom J, Chi J, Chen JH. Characterizing electronic health record usage patterns of inpatient medicine residents using event log data. PLoS One 2019 Feb 6;14(2):e0205379 [FREE Full text] [doi: 10.1371/journal.pone.0205379] [Medline: 30726208]
24. Soulakis ND, Carson M, Lee Y, Schneider D, Skeehan C, Scholtens D. Visualizing collaborative electronic health record usage for hospitalized patients with heart failure. J Am Med Inform Assoc 2015 Mar;22(2):299-311 [FREE Full text] [doi: 10.1093/jamia/ocu017] [Medline: 25710558]
25. Chen Y, Lorenzi N, Sandberg W, Wolgast K, Malin B. Identifying collaborative care teams through electronic medical record utilization patterns. J Am Med Inform Assoc 2017 Apr 01;24(e1):e111-e120 [FREE Full text] [doi: 10.1093/jamia/ocw124] [Medline: 27570217]
26. Chen Y, Patel M, McNaughton C, Malin B. Interaction patterns of trauma providers are associated with length of stay. J Am Med Inform Assoc 2018 Jul 01;25(7):790-799 [FREE Full text] [doi: 10.1093/jamia/ocy009] [Medline: 29481625]
27. Chen Y, Yan C, Patel MB. Network Analysis Subtleties in ICU Structures and Outcomes. Am J Respir Crit Care Med 2020 Dec 01;202(11):1606-1607 [FREE Full text] [doi: 10.1164/rccm.202008-3114LE] [Medline: 32931298]

28. Chen Y, Lehmann CU, Hatch LD, Schremp E, Malin BA, France DJ. Modeling Care Team Structures in the Neonatal Intensive Care Unit through Network Analysis of EHR Audit Logs. Methods Inf Med 2019 Nov;58(4-05):109-123 [FREE Full text] [doi: 10.1055/s-0040-1702237] [Medline: 32170716]
29. Yan C, Zhang X, Gao C, Wilfong E, Casey J, France D, et al. Collaboration Structures in COVID-19 Critical Care: Retrospective Network Analysis Study. JMIR Hum Factors 2021 Mar 08;8(1):e25724 [FREE Full text] [doi: 10.2196/25724] [Medline: 33621187]
30. Chen Y, Nyemba S, Malin B. Auditing medical records accesses via healthcare interaction networks. AMIA Annu Symp Proc 2012;2012:93-102 [FREE Full text] [Medline: 23304277]
31. Chen Y, Kho AN, Liebovitz D, Ivory C, Osmundson S, Bian J, et al. Learning bundled care opportunities from electronic medical records. J Biomed Inform 2018 Jan;77:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2017.11.014] [Medline: 29174994]
32. Epic UserWeb. URL: http://userweb.epic.com [accessed 2021-08-29]
33. Adelman JS, Applebaum JR, Schechter CB, Berger MA, Reissman SH, Thota R, et al. Effect of Restriction of the Number of Concurrently Open Records in an Electronic Health Record on Wrong-Patient Order Errors: A Randomized Clinical Trial. JAMA 2019 May 14;321(18):1780-1787 [FREE Full text] [doi: 10.1001/jama.2019.3698] [Medline: 31087021]
34. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. 2011 Presented at: 31st International Conference on Distributed Computing Systems Workshops; June 20-24, 2011; Minneapolis, MN, USA p. 166-171. [doi: 10.1109/icdcsw.2011.20]
35. Chen B, Alrifai W, Gao C, Jones B, Novak L, Lorenzi N, et al. Mining tasks and task characteristics from electronic health record audit logs with unsupervised machine learning. J Am Med Inform Assoc 2021 Jun 12;28(6):1168-1177. [doi: 10.1093/jamia/ocaa338] [Medline: 33576432]
36. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. 2009 Presented at: Proceedings of the International AAAI Conference on Web and Social Media; March 19, 2009; San Jose, CA, USA p. 361-362.
37. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O'Neal L, REDCap Consortium. The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019 Jul;95:103208 [FREE Full text] [doi: 10.1016/j.jbi.2019.103208] [Medline: 31078660]

Abbreviations

ANOVA: analysis of variance
ECMO: extracorporeal membrane oxygenation
EHR: electronic health record
HCO: health care organization
HCW: health care worker
NICU: neonatal intensive care unit
PCA: principal component analysis
t-SNE: t-distributed stochastic neighbor embedding
WSS: within-cluster sum of squared errors

Edited by G Eysenbach; submitted 21.03.21; peer-reviewed by H Spallek, H Salmasian, A Hughes; comments to author 13.04.21; revised version received 23.05.21; accepted 02.08.21; published 03.09.21.

Please cite as:
Li P, Chen B, Rhodes E, Slagle J, Alrifai MW, France D, Chen Y
Measuring Collaboration Through Concurrent Electronic Health Record Usage: Network Analysis Study
JMIR Med Inform 2021;9(9):e28998
URL: https://medinform.jmir.org/2021/9/e28998
doi:10.2196/28998
PMID:34477566

Review

Models Predicting Hospital Admission of Adult Patients Utilizing Prehospital Data: Systematic Review Using PROBAST and CHARMS

Ann Corneille Monahan[1], MSHI, PhD; Sue S Feldman[2], RN, MEd, PhD

[1]Department of Epidemiology & Public Health, School of Public Health, University College Cork, Cork, Ireland

[2]Department of Health Services Administration, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Ann Corneille Monahan, MSHI, PhD
Department of Epidemiology & Public Health
School of Public Health
University College Cork
College Road
Cork, T12 K8AF
Ireland
Phone: 353 21 420 5860
Email: monahanannc@gmail.com

Abstract

Background: Emergency department boarding and hospital exit block are primary causes of emergency department crowding and have been conclusively associated with poor patient outcomes and major threats to patient safety. Boarding occurs when a patient is delayed or blocked from transitioning out of the emergency department because of dysfunctional transition or bed assignment processes. Predictive models for estimating the probability of an occurrence of this type could be useful in reducing or preventing emergency department boarding and hospital exit block, to reduce emergency department crowding.

Objective: The aim of this study was to identify and appraise the predictive performance, predictor utility, model application, and model utility of hospital admission prediction models that utilized prehospital, adult patient data and aimed to address emergency department crowding.

Methods: We searched multiple databases for studies, from inception to September 30, 2019, that evaluated models predicting adult patients' imminent hospital admission, with prehospital patient data and regression analysis. We used PROBAST (Prediction Model Risk of Bias Assessment Tool) and CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modeling Studies) to critically assess studies.

Results: Potential biases were found in most studies, which suggested that each model's predictive performance required further investigation. We found that select prehospital patient data contribute to the identification of patients requiring hospital admission. Biomarker predictors may add superior value and advantages to models. It is, however, important to note that no models had been integrated with an information system or workflow, operated independently as electronic devices, or operated in real time within the care environment. Several models could be used at the site-of-care in real time without digital devices, which would make them suitable for low-technology or no-electricity environments.

Conclusions: There is incredible potential for prehospital admission prediction models to improve patient care and hospital operations. Patient data can be utilized to act as predictors and as data-driven, actionable tools to identify patients likely to require imminent hospital admission and reduce patient boarding and crowding in emergency departments. Prediction models can be used to justify earlier patient admission and care, to lower morbidity and mortality, and models that utilize biomarker predictors offer additional advantages.

(JMIR Med Inform 2021;9(9):e30022) doi:10.2196/30022

KEYWORDS

emergency services; hospital; decision support techniques; patient-specific modeling; crowding; boarding; exit block; systematic review; PROBAST; CHARMS; predictive model; medical informatics; health services research; prehospital assessment; process improvement; management information system; predict admission; emergency department

Introduction

Background

The delivery of timely quality care in emergency departments has become increasingly challenging due to crowding [1,2]. Emergency department crowding is an international problem [3-5] that has been of continuing concern for the last two decades and is expected to become more problematic with population growth and an aging population whose life expectancy is increasing. The magnitude of the crowding problem has been demonstrated by decades of research into emergency department efficiency interventions that aimed to reduce crowding by improving throughput and processes, such as triage, diagnosis, and treatment, that affect the flow of care [6,7]. However, these measures primarily promoted efficiency in portions of the emergency department care continuum and had little effect in reducing crowding, because they did not address the source of the problem at a system level [8].

Rigorous analysis suggests that exit block and emergency department boarding are the main causes of emergency department crowding [6,9-12]. Boarding is the retention of patients who have already been admitted to the hospital in the emergency department because they await assignment to an inpatient hospital bed [5]. Exit block is the delay that occurs when patients cannot be transitioned into the hospital for admission or discharged (home, rehabilitation, etc) in a timely manner [5,8]. Exit block results in emergency department boarding and is a system issue [8,13]. Both boarding and the resulting overcrowding have been conclusively associated with poor patient outcomes and threats to patient safety [5,14-17].

Predictive Modeling

Predictive modeling that can be used to address emergency department crowding is an emerging field of study. Predictive modeling is used to anticipate which factors will bring about a particular outcome [18]. In health care, models use specific data to estimate the probability that a condition or disease is already present (a diagnostic model) or the probability that an outcome will occur in the future (a prognostic model) [18]. Recent studies [19-28] of models utilizing these techniques estimate patient risk for health conditions and patient–provider encounters (eg, suicide attempts or intentional acts of self-harm) [19], acute kidney injury (ie, sudden kidney failure or damage) [20], hospital readmissions (ie, readmission to a hospital within 30 days of discharge, regardless of cause) [23,24,26,27], and perioperative mortality (ie, deaths within 30 days of surgery) [21], emergency department return visits (ie, return emergency department visits within 72 hours for any reason) [28], return visits after hospital discharge (ie, return emergency department visits within 30 days of hospital discharge for any reason) [25], and emergency department crowding or demand (ie, the availability of space for patients relative to the volume of patients that need to be seen) [22]) to improve health care delivery and patient outcomes. A subsection of this area of study focuses on predicting which emergency department patients are likely to require imminent hospital admission. This area of research is important because of its direct and immediate potential to lower patient morbidity and mortality by helping emergency department patients receive care earlier in the emergency department care continuum.

While more prediction models have been developed in recent years [18], external validation studies of published prediction models have not kept pace [29]. There is often no consensus about the best, most effective model for a particular purpose, leaving providers and policy makers unable to choose a model with confidence. In the case of hospital admission prediction, most models have not been externally validated or tested in a live emergency department environment. Furthermore, systematic reviews have received scrutiny for their lack of rigor [30-32]. Hence, a rigorous systematic review of studies of admission prediction models is needed to synthesize findings that researchers and decision-makers can rely on with confidence to address localized emergency department boarding, crowding, and exit block, as well as system-wide implications.

Systematic Review Validation

Rigorous systematic reviews follow accepted approaches. PROBAST (Prediction Model Risk of Bias Assessment Tool) [33] can be used to identify potential sources of bias in individual prediction model studies, and CHARMS (Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies) [34] can also be used to identify potential sources of bias, organize information, and identify relevant information used to evaluate the prediction modeling studies. While the systematic review of clinical trials is generally a well-established field, the fields of health care prediction modeling and systematic review of such studies are not as well established, despite growth in these fields. For example, a search of Google Scholar for “systematic review” AND “prediction” AND “healthcare” demonstrated an increase of 410% in publications between decades (from n=45,900 in 2000-2010 to n=234,000 in 2010-2020). As the number of prediction modeling publications continue to grow, the need exists to apply the same rigor to systematic reviews of health care–related prediction modeling as that which has been applied to clinical trial and other types of systematic reviews through the use of tools, such as PROBAST and CHARMS, to facilitate quality assessment for individual prediction model studies using standardized guidelines [30,33]. Only two systematic reviews [35,36] that have focused on increasing overall throughput by decreasing emergency department boarding and systemic exit block in health systems applied the rigorous PROBAST and CHARMS methodologies, with both reporting a high degree of bias in the studies that they examined.

Logistic Regression for Systematic Reviews

Logistic regression is a technique for understanding the relationships between predictor variables and outcomes and is one of the most commonly used methods for forecasting [37]. There are a variety of techniques that can be used to model data; each is designed to accommodate types of data, number of predictors, and study aims, and each has advantages and disadvantages. Logistic regression is only used for data with a binary outcome and multiple predictors and accommodates predictors of multiple data types, such as continuous and categorical data; therefore, data types do not need to be modified, which can introduce potential bias. Logistic regression produces a mathematical form—a weighted combination of variables that predict the outcome variable [37].

We aimed to better understanding predictive modeling's role in addressing the emergency department crowding problem by examining model predictive performance, the utility of the contribution of prehospital patient data to model prediction, applications of models, and the utility of models.

Methods

Study Design

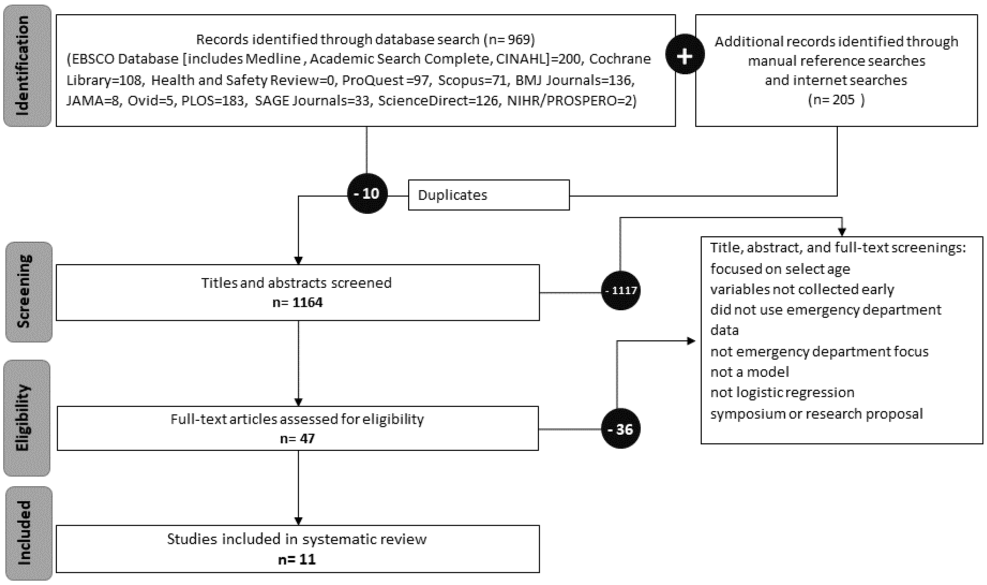
We applied PROBAST and CHARMS to rigorously assess studies of models designed to predict adult patient imminent hospital admission using prehospital patient data collected early in the emergency department visit or during ambulance transport to the emergency department. We searched databases for papers published from inception through September 30, 2019. Data were organized and analyzed in Excel (version 2016, Microsoft Inc). This study did not require institutional review board authorization.

Data Sources and Search Strategy

We reviewed database content descriptions for 99 health science, public health, and medical databases to determine their relevance to our topic of interest, and 13 databases were found to be relevant: EBSCO Database (includes Medline database and Academic Search Complete database), CINAHL Plus with Full Text, Cochrane Library, Health and Safety Review, ProQuest Central, Scopus, BMJ Journals, JAMA, Journals at Ovid, PLOS, SAGE Journals, ScienceDirect, and NIHR/PROSPERO.

The *Title, abstract, or keyword* option was used with the following search string: "model or strategy and hospital* and predict* or risk." (Asterisks were used to capture hospital, hospitalization, hospitalisation, hospitalized, hospitalized and predict, predicts, predicted, predictor, predictive.) If no results were initially produced, the search was expanded by removing all filters and searching for the terms anywhere in the document. Sources that did not allow for truncation were searched multiple times with multiple word combinations. Additionally, the internet was searched with the following combined terms: "model predict hospital admission," "risk of hospital admission," "hospital admission model," "admission risk," "emergency model," and "hospital admission." Reference lists were also reviewed (Figure 1).

Figure 1. Search flow diagram of included studies.


Identification
Records identified through database search (n= 969)
(EBSCO Database [includes Medline , Academic Search Complete, CINAHL]=200, Cochrane Library=108, Health and Safety Review=0, ProQuest =97, Scopus=71, BMJ Journals=136, JAMA=8, Ovid=5, PLOS=183, SAGE Journals=33, ScienceDirect=126, NIHR/PROSPERO=2)
+
Additional records identified through manual reference searches and internet searches (n= 205)
- 10
Duplicates
Screening
Titles and abstracts screened n= 1164
- 1117
Title, abstract, and full-text screenings:
focused on select age
variables not collected early
did not use emergency department data
not emergency department focus
not a model
not logistic regression
symposium or research proposal
Eligibility
Full-text articles assessed for eligibility n= 47
- 36
Included
Studies included in systematic review n= 11

Inclusion and Exclusion Criteria

We included full-text peer-reviewed English-language studies that evaluated strategies or models using prehospital patient data to predict imminent hospital admission of primarily adult general medicine patients with regression.

Studies in which the setting was not an emergency department, data were not collected early in the emergency department visit, or either models or logistic regression were not used and that focused on pediatric (<16 years of age), psychiatric, or specific health conditions were excluded.

Data Quality Assessment

We used PROBAST to assess risk of bias for each study. Shortcomings in a study's design, conduct, or analysis can cause systematic errors that result in flawed or distorted results and hamper internal validity [18]. Assessment of the quality of studies, including risk of bias and model applicability to the target settings and populations, is an essential component of systematic reviews and their evidence synthesis. The first step in applying PROBAST was the identification of a clear and focused review question about the intended use of the model, targeted participants, predictors used in the modeling, and predicted outcome [33]. The second step was the identification and assessment of potential sources of bias in 4 domains (participants, predictors, outcomes, analysis). Key qualities assessed for each study included the appropriateness of the data source, whether predictors were similarly measured and defined, whether outcomes were measured similarly for all participants, and whether missing data were appropriately handled and reported.

Data Extraction and Data Synthesis

We used CHARMS to identify key items in 11 domains (eg, source of data, sample size, model development, model performance, results) in individual studies (and in their PROBAST reports) in order to evaluate potential sources of bias and issues that may affect the applicability of results in relation to the intended use of the model. Key information was organized by relevant domains (Multimedia Appendix 1).

Results

General

Searches produced 1164 citations, from which 47 were selected for full review; 11 studies met inclusion criteria. Each model was critically assessed with PROBAST (Multimedia Appendix 2) and CHARMS.

CHARMS Study Characteristics

Data Source, Participants, and Outcome CHARMS Domains 1, 2, and 3

Of the 11 studies, 3 used a prospective observational cohort [38-40], and the remaining 8 used a retroactive observational cohort [22,41-47]. There was good diversity, in terms of the countries in which studies took place (South Africa [38], Scotland [41], the United States [22,42,44,45], the Netherlands [40,43], Australia [39], and Singapore [46,47]). Sampling ranged from 14 days [40] to 10 years [46], with most study durations between 3 and 27 months [38,39,41-43,45,47]. Two studies were 2 months in length [22,44].

Most studies utilized clinical and administrative patient information collected early in the emergency visit [22,38-43,46,47]; 2 studies used data collected during ambulance transport to the emergency department [44,45]. Additionally, all studies evaluated 1 or more models' abilities to predict patient imminent need for hospital admission and defined outcome event by patient final disposition, and measured outcome by patient hospital admission or discharge from the emergency department. Furthermore, all studies corresponded to the outcome definition of the systematic review question, which reduced the potential for bias from different outcome definitions and measurement methods that can lead to differences in study results and would be a source of heterogeneity across studies [34].

Candidate Predictors CHARMS Domain 4

Candidate predictors included all predictors investigated in a given study for predictive performance and not the finalized predictors included in model analysis. Candidate predictors ranged from 5 to 14 per study (Multimedia Appendix 3): under 10 predictors [22,38,39,45], over 10 predictors [40,41,43,44,46,47], and did not report [42]. Overall, 52 candidate predictors had been evaluated, and 34 predictors were retained in models (across all studies).

Sample Size CHARMS Domain 5

Consideration of sample size is important to ensure adequate numbers of data events are collected to achieve meaningful results. Sample sizes ranged from 401 to 864,246. None reported sample size calculation, estimation, or rationale. One study [40] did, however, perform a sample size calculation for its validation. All studies described efforts to avoid overfitting, which included model comparison to validation models [22,38,40,41,43,44,46,47], model comparison to multiple site outcomes [45], model comparison to published models [42], and model comparison to triage nurse prediction of patient final disposition [39]. Overfitting describes when findings in the development sample do not exist in the relevant population resulting in a model that too closely fits the development data set and produces findings that are not reproducible [37]. Overfitting is a primary concern in prediction modeling development that can be mitigated by performing sample size estimates during study design [34].

Missing Data CHARMS Domain 6

Infrequently is value attributed to missing data in the missing state [48]; instead, the missing values are either imputed or disregarded completely [49,50]. Four studies described a process for handling missing data: 3 used multiple imputation [39,41,43], and 1 study reported "missing predictors were replaced with missing values" [42]; it was unknown whether this referred to blank (ie, missing) identifiers or whether missing values were imputed. Of the remaining 7 studies, 1 study reported 30% of data were missing and did not describe how missing data were handled (ie, whether the patient events were included or excluded) [38], and 6 studies did not mention missing data at all [22,40,44-47].

Model Development CHARMS Domain 7

Two studies also developed models using other techniques (gradient boosting and deep neural network [42], and naive Bayes [22]) in addition to models using logistic regression. Most studies selected predictors using univariate analysis [22,39,40,42,43,46,47], but 4 studies used multivariate modeling [38,41,44,45].

Model Performance CHARMS Domain 8

Model predictive performance was gauged via the percentage of patients actually admitted, the percentage of patients predicted to be admitted, and goodness of fit tests that assessed model discrimination and model calibration (Table 1).

Table 1. Model performance predicting patient hospital admission.

Reference	Model performance			
	Admission		Goodness of fit tests	
	Actual, n (%)	Predicted, %	Discrimination, AUROC[a] (95% CI)	Calibration[b]
Burch et al [38]	469 (59)	—[c]	—	—
Cameron et al [41]	—	—	0.88 (0.88-0.88)	—
Hong et al [42]	60,277 (29.7)	—	0.86(0.86-0.87)	—
Kim et al [39]	38,695 (38.6)	—	0.80 (0.80-0.80)	Performed, not reported
Kraaijvanger et al [40]	400 (31.7)	31.1	0.87 (0.85-0.89)	Reported to be good
Lucke et al [43]	2912 (27)	21.4	0.86 (0.85-0.87)	Reported to be good
Meisel et al [44]	132 (33)	32	0.80 (—)	Performed, not reported
Meisel et al [45]	440 (24.8)	39.8	0.83 (—)	—
Parker et al [46]	334,115 (38.7)	—	0.83 (0.82-0.83)	Reported to be good
Peck et al [22]	—	—	0.89 (—)	r^2=0.58 moderate to poor
Sun et al [47]	95,909 (30.2)	30	0.85 (0.85-0.85)	Reported to be good

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

[b] Studies used several formulas to evaluate calibration, to include Hosmer-Lemeshow, threshold probability, and r^2.

[c] Not reported.

Discrimination is a model's ability to distinguish between patients who do and do not experience the outcome of interest and is most commonly assessed with the area under the receiver operating characteristics (AUROC) [51]. The AUROC represents the performance of a classification model that has a categorical outcome, producing a score representing a proportion of times the model correctly discriminated between groups, for example, those at high risk and low risk. The higher the AUROC, the better the model discriminates between the 2 groups (0.5-0.6 represents not better than chance, 0.6-0.7 represents poor, 0.7-0.8 represents fair, 0.8-0.9 represents good, and 0.9-1.0 represents excellent discrimination [52]). Eight studies reported good discrimination [22,40,47], 2 reported fair discrimination [39,53], and 1 study did not report any performance measurement [38].

Calibration is the extent to which model predicted risk compares to observed outcomes (ie, difference between rates of observed events and predicted events for groups [54]. Calibration is usually reported graphically by plotting observed against predicted event rates [55] and is commonly measured with the Hosmer-Lemeshow statistical test for binary categorical outcomes [54]. Most studies that measured calibration statistically, reported good agreement between predicted and observed hospital admission. Seven models evaluated calibration using Hosmer-Lemeshow [44,47,43,39], threshold probability of admission [46], or R^2 [22], 1 did not report which statistic was used [40], and 2 of these 7 studies did not report results [39,44]. Four studies did not measure calibration [38,41,42,45].

Model Evaluation: Domain 9

Utility of predictive models depends on their external validation—performance evaluation on an independent data set. External validation took a variety of forms: different settings with different samples [40], same locations with different samples [43,45,46], and nurse opinion on likely patient admission [22,39]. Five models were internally validated [38,41,42,44,47].

Model Results: Domain 10

Predictive accuracy and precision drive model performance and the extent to which it can estimate the probability of individual patient outcomes, as well as model suitability for clinical and administrative uses.

The models in the 11 studies were not operational (no apps developed and no integration with information systems or workflow) and were not tested in environments in which they would be used, which compromised the evaluation of model feasibility. Operational models would identify patients likely to require hospital admission; thus, there is a great amount of utility and potential for models to improve patient care and hospital operations, including by reducing hospital exit block, emergency department boarding, and ultimately emergency department crowding.

Interpretation and Discussion: Domain 11

The utility of select prehospital patient data to act as predictors and as data-driven, actionable tools to identify patients requiring hospital admission was shown. The models utilizing biomarker predictors (eg, blood pressure, heart rate) [38,43,45] may provide advantages due to standardized definition, measurement, and interpretation of these biomarker measures. Models that use only biomarker predictors may be widely applicable and robust, and their results may be generalizable to populations and environments. Models that did not include patient history variables (eg, chronic conditions, number of prior emergency department visits) [22,38,40,47] may have greater applicability because the model does not rely on the availability of medical record information or patient reports. The predictors in these models—prehospital patient data collected early in the emergency department visit or during ambulance transport—are not the only options for predicting patient admission but are likely the best options for making timely predictions using data collected in the early stages of an urgent care visit.

AUROC values suggested fair to good ability to distinguish between outcome groups (admitted, not admitted), and thus, to predict patient imminent need for hospital admission. Likewise, the utility of the variables as predictors for the identification of patients likely to require imminent hospital admission was shown.

Risk of Bias Assessment

Data transformation can increase risk of bias by satisfying assumptions without changing the scale of representation [56]. Five studies did not transform raw data [38,44-47]. On the other hand, 6 studies transformed predictors, such as, by categorizing continuous variables and dichotomizing continuous variables [22,39-43].

Evaluation of heterogeneous predictors across studies introduces bias if they are treated as identical. In 2 studies, bias was low, because standardized, frequently calibrated equipment was used to measure predictors (eg, blood pressure, laboratory analysis, etc), which produces measurements that are comparable across studies, required no manipulation (eg, dichotomized, categorized), and offer more likelihood of retaining reliability when applied to new populations [38,43]. Age has been shown to inject bias, for example, the same model can appear to perform better when applied to a sample with a wide age range than when applied to a sample with a narrow age range [57]. Nine models included age [22,39-41,43-47], with only 2 studies indicating age >60 years [44,45].

Estimating sample size during study design minimizes model overfitting and includes calculating events-per-variable. Events-per-variable, generally, is poorly reported in prediction model studies [34] and was not reported in any of the included studies. However, events-per-variable can be calculated from other study information to aid assessment of study quality. The appropriateness of most studies' sample size could be evaluated by calculating study events-per-variable, the number of data events needed per predictor variable to achieve meaningful results [37]. This ratio was calculated using study limiting sample size, the portion of outcome events (admitted or not admitted) that is smaller [37]. The focus is on the smaller portion of outcome events, because the total sample size is not directly relevant in binary models [37]. The limiting sample size is divided by the number of candidate predictors to produce the limiting events-per-variable ratio.

In 10 studies [22,39-47], the limiting sample size was the number of admitted patients, but in 1 study [38] the limiting sample size was the number of patients who were not admitted (ie, more patients were admitted than discharged). Limiting events-per-variable could not be calculated for 3 models because either the proportion of admitted patients or the number of candidate predictors was not reported [22,41,42]. The limiting sample size range of studies was 132.3 to 334,115, producing a limiting events-per-variable range of 9 to 30,374. The limiting events-per-variable was sufficient in most studies to obtain meaningful results and avoid bias from an overfitted model. However, at 9 events-per-variable, 1 model [44] was below the recommended 10 to 15 events-per-variable [42,58,59] and was in jeopardy of bias.

Missing data handling can inject bias. To mitigate against bias with imputation, 3 studies used multiple imputation [39,41,43], substituting missing observations with plausible estimated values derived from analysis of available data, which is the preferred method for handling missing data in prediction research [34,60]. One study [42] reported replacing missing values but did not disclose how these missing values were placed, and the remaining 7 studies did not describe the handling of missing data [22,38,40,44-47], which suggested there was an element of risk of bias. Data are usually not missing at random and instead are related to other observed participant data and, as a consequence, participants with complete data are different from those with incomplete data [34,61].

Per PROBAST definition, a model that is internally validated is a development-only study—not a development and validation study. A model must be externally validated to be considered a development and validation study. While 6 of the models were externally validated [22,39,40,43,45,46], 2 studies used nurses' opinions [22,39] and were not validated with data.

Inclusion of false predictors increases the likelihood of model overfitting because the model corresponds too closely to its derivation data set and fails to fit other relevant data sets or predict future observations reliably [62], resulting in overly optimistic predictions of model performance for new data sets [34]. In univariate analysis, each predictor is tested individually for its association with the outcome, and the most statistically significant predictors are included in the model. However, univariate analysis is not the preferred method because it commonly introduces selection bias when predictors selected for model inclusion have a large but false association with the outcome [18,63]. In small samples, predictors could initially show no association with outcome, but after adjustment for other predictors, may show association with the outcome [34]. Conversely, multivariate modeling is preferred for predictor selection because there is no selection bias since all predictors are prespecified. Only 4 of the models used multivariate modeling for predictor selection [38,41,44,45], and the remaining models used univariate analysis [22,39,40,42,43,46,47].

Discussion

Principal Findings

This study showed the utility of select, prehospital patient data to act as predictors to model identification of patients likely to require hospital admission and that models produced information that could be used to improve patient care and hospital operations. Ten studies reported model discrimination with AUROC: 8 studies reported values [22,40-43,45-47] that suggest good ability to distinguish between outcome groups (admitted, not admitted), and thus, to predict patients' imminent need for hospital admission. An example of model application for patients who are predicted to require admission is earlier bed request giving managers more time to secure a patient bed. This forewarning could result in operations procedures to decrease exit block and increase patient flow out of the emergency department [13].

Potential sources of bias that may cause flawed or distorted model predictions were found in every model, for example, from minor (not reporting handling of missing values [38,39,43,44,47], univariate predictor selection [39,47]) to potentially damaging (dichotomized continuous variables [22,41,43], low events-per-variable [44], no external validation [38,41,42,44,47]), which suggest that study reports of models' abilities to predict outcomes have the potential to be flawed. This is consistent with other evaluations of prediction modeling studies [34], including evaluations applying CHARMS and PROBAST in the emergency department setting [35,36].

Overall, model performances were reportedly good, with most models showing good ability to discriminate between patients who do and do not require imminent hospital admission [22,40-43,45-47], and almost half reporting good calibration to detect differences between observed and predicted admission rates [40,43,46,47]. Although several studies did not measure calibration [38,41,42,45], the remainder did [22,39,40,43,44,46,47]. However, all [38-47] but 1 study [22] poorly reported its measurement. Findings of neglected calibration measures, with an overreliance on discrimination measures, are consistent with those of other reports [34]. Assessing and reporting discrimination and calibration are important in prediction model evaluation. No models were found to have operated through an app, and none had been integrated with an information system. However, to function as intended, most models required development of an electronic app to receive patient data, operate the algorithm, and produce results. Most also required app integration with an information system to produce real-time admission prediction. Studies also did not describe a process to achieve app development or system integration.

Biomarker predictors may contribute superior value and advantage to a model due to their lack of variability in definition, measurement, and interpretation, and freedom from the confines of patient histories, resulting in a widely applicability.

The quantity of candidate predictors demonstrated the breadth of potential influences on patients' imminent need for hospital admission. However, the number of predictors across studies did not reflect the quantity accurately because, across studies, multiple names were used for the same predictor—identically named predictors were defined differently, data collection and evaluation varied, and predictors composed of multiple variables were not specified

Models have the potential to facilitate hospital admission, subsequently reducing or ending hospital exit block, emergency department boarding, and emergency department crowding but none had been implemented or tested.

To develop models with the most potential, future investigations must address deficiencies, avoid risk of bias in model design and investigation, verify the utility of biomarker predictors and the most useful predictor combination, evaluate real-time utility of admission prediction on hospital operations, compare performance of technology enabled versus intuition, and verify longitudinal model impact on patient care and hospital operations.

Limitations

Although the findings of this review are valuable and add to the current literature on artificial intelligence models in the emergency department setting, this study has several limitations. First, this was a critique of the methodologies used in the models; we did not consider the feasibility of the models examined. Second, the selection of studies and PROBAST assessments were performed by one researcher, with a second researcher providing oversight. The use of multiple researchers would have ensured intercoder reliability and mitigated systematic errors. Additionally, only studies in English and conducted with emergency department setting data were included. That being said, this study closely adhered to the CHARMS methodology for study evaluation.

Comparison With Prior Work

We applied both CHARMS and PROBAST to studies that used logistic regression and data from emergency department settings. Our findings are consistent with those of previous systematic reviews [35,36,64,65] that applied PROBAST and CHARMS methodologies to evaluate health care prediction models, in terms of risk of bias. We attempted to be focused and provide depth of analysis by identifying and appraising hospital admission prediction models that utilized prehospital patient data in a defined setting (emergency department). Four healthcare prediction model studies were reviewed for their use of PROBAST and CHARMS methodologies. However, while 2 [35,36] were set in the emergency department, evaluation variables and outcome of interest differed for all 4 studies [35,36,64,65].

Acknowledgments

ACM conceived the study design, conducted the literature review and analysis, and contributed to writing the manuscript. SSF provided oversight on the study design and literature analysis and contributed to writing the manuscript.

Conflicts of Interest

SSF receives consultancy fees from Guideway Cares (which are not in relation to this work).

Multimedia Appendix 1
Study characteristics by CHARMS domains.
[DOCX File , 34 KB - medinform_v9i9e30022_app1.docx]

Multimedia Appendix 2
Completed PROBAST.
[DOCX File , 53 KB - medinform_v9i9e30022_app2.docx]

Multimedia Appendix 3
Predictors evaluated by each study.
[DOCX File , 29 KB - medinform_v9i9e30022_app3.docx]

References

1. Sinclair D. Emergency department overcrowding - implications for paediatric emergency medicine. Paediatr Child Health 2007 Jul;12(6):491-494 [FREE Full text] [doi: 10.1093/pch/12.6.491] [Medline: 19030415]
2. American College of Emergency Physicians (ACEP). Crowding. policy statement. Ann Emerg Med 2013 Jun;61(6):726-727. [doi: 10.1016/j.annemergmed.2013.03.037] [Medline: 23684339]
3. Di Somma S, Paladino L, Vaughan L, Lalle I, Magrini L, Magnanti M. Overcrowding in emergency department: an international issue. Intern Emerg Med 2014 Dec 2;10(2):171-175. [doi: 10.1007/s11739-014-1154-8] [Medline: 25446540]
4. Higginson I, Boyle A. What should we do about crowding in emergency departments? Br J Hosp Med (Lond) 2018 Sep 02;79(9):500-503 [FREE Full text] [doi: 10.12968/hmed.2018.79.9.500] [Medline: 30188202]
5. Richards JR, van der Linden C, Derlet RW. Providing Care in Emergency Department Hallways: Demands, Dangers, and Deaths. Adv Emerg Med 2014 Dec 25;2014:1-7 [FREE Full text] [doi: 10.1155/2014/495219]
6. Stead LG, Jain A, Decker WW. Emergency department over-crowding: a global perspective. Int J Emerg Med 2009 Sep 30;2(3):133-134 [FREE Full text] [doi: 10.1007/s12245-009-0131-x] [Medline: 20157461]
7. Kauppila T, Seppänen K, Mattila J, Kaartinen J. The effect on the patient flow in a local health care after implementing reverse triage in a primary care emergency department: a longitudinal follow-up study. Scand J Prim Health Care 2017 Jun 08;35(2):214-220 [FREE Full text] [doi: 10.1080/02813432.2017.1333320] [Medline: 28593802]
8. Henderson K, Boyle A. Exit block in the emergency department: recognition and consequences. Br J Hosp Med (Lond) 2014 Nov 02;75(11):623-626 [FREE Full text] [doi: 10.12968/hmed.2014.75.11.623] [Medline: 25383431]
9. Higginson I. Emergency department crowding. Emerg Med J 2012 Jun 04;29(6):437-443. [doi: 10.1136/emermed-2011-200532] [Medline: 22223713]
10. Institute of medicine. Hospital-Based Emergency Care: At the Breaking Point. Washington, DC: National Academies Press; 2006.
11. Mason S, Knowles E, Boyle A. Exit block in emergency departments: a rapid evidence review. Emerg Med J 2017 Jan 27;34(1):46-51. [doi: 10.1136/emermed-2015-205201] [Medline: 27789568]
12. Scott I, Sullivan C, Staib A, Bell A. Deconstructing the 4-h rule for access to emergency care and putting patients first. Aust Health Rev 2018;42(6):698. [doi: 10.1071/ah17083] [Medline: 29032791]
13. Orewa G, Feldman SS, Hearld KR, Kennedy KC, Hall AG. Using accountable care teams to improve timely discharge: a pilot study. Qual Manag Health Care 2021 Aug 03:1. [doi: 10.1097/QMH.0000000000000320] [Medline: 34354033]
14. Carter EJ, Pouch SM, Larson EL. The relationship between emergency department crowding and patient outcomes: a systematic review. J Nurs Scholarsh 2014 Mar 19;46(2):106-115 [FREE Full text] [doi: 10.1111/jnu.12055] [Medline: 24354886]
15. Reznek MA, Murray E, Youngren MN, Durham NT, Michael SS. Door-to-imaging time for acute stroke patients is adversely affected by emergency department crowding. Stroke 2017 Jan;48(1):49-54. [doi: 10.1161/strokeaha.116.015131] [Medline: 27856953]
16. Odom N, Babb M, Velez L, Cockerham Z. Stud Health Technol Inform 2018;250:178-181. [Medline: 29857424]
17. Eriksson CO, Stoner RC, Eden KB, Newgard CD, Guise J. The association between hospital capacity strain and inpatient outcomes in highly developed countries: a systematic review. J Gen Intern Med 2017 Jun 15;32(6):686-696 [FREE Full text] [doi: 10.1007/s11606-016-3936-3] [Medline: 27981468]
18. Moons KG, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. Ann Intern Med 2019 Jan 01;170(1):w1-w33. [doi: 10.7326/m18-1377] [Medline: 30596876]
19. Simon GE, Johnson E, Lawrence JM, Rossom RC, Ahmedani B, Lynch FL, et al. PPredicting suicide attempts and suicide deaths following outpatient visits using electronic health records. Am J Psychiatry 2018 Oct 01;175(10):951-960 [FREE Full text] [doi: 10.1176/appi.ajp.2018.17101167] [Medline: 29792051]

20. Mohamadlou H, Lynn-Palevsky A, Barton C, Chettipally U, Shieh L, Calvert J, et al. Prediction of acute kidney injury with a machine learning algorithm using electronic health record data. Can J Kidney Health Dis 2018 Jun 08;5:2054358118776326 [FREE Full text] [doi: 10.1177/2054358118776326] [Medline: 30094049]
21. Garcea G, Ganga R, Neal CP, Ong SL, Dennison AR, Berry DP. Preoperative early warning scores can predict in-hospital mortality and critical care admission following emergency surgery. J Surg Res 2010 Apr;159(2):729-734. [doi: 10.1016/j.jss.2008.08.013] [Medline: 19181337]
22. Peck J, Benneyan J, Nightingale D, Gaehde S. Predicting emergency department inpatient admissions to improve same-day patient flow. Acad Emerg Med 2012 Sep;19(9):E1045-E1054 [FREE Full text] [doi: 10.1111/j.1553-2712.2012.01435.x] [Medline: 22978731]
23. Allaudeen N, Vidyarthi A, Maselli J, Auerbach A. Redefining readmission risk factors for general medicine patients. J Hosp Med 2011 Feb 12;6(2):54-60. [doi: 10.1002/jhm.805] [Medline: 20945293]
24. Mudge AM, Kasper K, Clair A, Redfern H, Bell JJ, Barras MA, et al. Recurrent readmissions in medical patients: a prospective study. J Hosp Med 2011 Feb 12;6(2):61-67. [doi: 10.1002/jhm.811] [Medline: 20945294]
25. Li C, Chang H, Wang H, Bai Y. Diabetes, functional ability, and self-rated health independently predict hospital admission within one year among older adults: a population based cohort study. Arch Gerontol Geriatr 2011 Mar;52(2):147-152. [doi: 10.1016/j.archger.2010.03.004] [Medline: 20338646]
26. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, et al. Risk prediction models for hospital readmission: a systematic review. JAMA 2011 Oct 19;306(15):1688-1698 [FREE Full text] [doi: 10.1001/jama.2011.1515] [Medline: 22009101]
27. Nguyen OK, Makam AN, Clark C, Zhang S, Das SR, Halm EA. Predicting 30 - Day Hospital Readmissions in Acute Myocardial Infarction: The AMI “READMITS” (renal function, elevated brain natriuretic peptide, age, diabetes mellitus nonmale sex intervention with timely percutaneous coronary intervention, and low systolic blood pressure) score. JAHA 2018 Apr 17;7(8):e008882. [doi: 10.1161/jaha.118.008882]
28. Bergese I, Frigerio S, Clari M, Castagno E, De Clemente A, Ponticelli E, et al. An innovative model to predict pediatric emergency department return visits. Pediatr Emerg Care 2019 Mar;35(3):231-236. [doi: 10.1097/PEC.0000000000000910] [Medline: 27741066]
29. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016 Jan;69:245-247 [FREE Full text] [doi: 10.1016/j.jclinepi.2015.04.005] [Medline: 25981519]
30. Brackett A, Batten J. Ensuring the rigor in systematic reviews: Part 1, the overview. Heart Lung 2020;49(5):660-661. [doi: 10.1016/j.hrtlng.2020.03.015] [Medline: 32532424]
31. Ioannidis JP. he mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q 2016 Sep 13;94(3):485-514 [FREE Full text] [doi: 10.1111/1468-0009.12210] [Medline: 27620683]
32. McIntosh GS, Steenstra I, Hogg-Johnson S, Carter T, Hall H. Lack of prognostic model validation in low back pain prediction studies: a systematic review. Clin J Pain 2018 Aug;34(8):748-754. [doi: 10.1097/AJP.0000000000000591] [Medline: 29406366]
33. Wolff RF, Moons KG, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med 2019 Jan 01;170(1):51. [doi: 10.7326/m18-1376] [Medline: 30596875]
34. Moons KM, de Groot JA, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Med 2014 Oct;11(10):e1001744 [FREE Full text] [doi: 10.1371/journal.pmed.1001744] [Medline: 25314315]
35. Miles J, Turner J, Jacques R, Williams J, Mason S. Using machine-learning risk prediction models to triage the acuity of undifferentiated patients entering the emergency care system: a systematic review. Diagn Progn Res 2020 Oct 02;4(1):16 [FREE Full text] [doi: 10.1186/s41512-020-00084-1] [Medline: 33024830]
36. Kareemi H, Vaillancourt C, Rosenberg H, Fournier K, Yadav K. Machine learning versus usual care for diagnostic and prognostic prediction in the emergency department: a systematic review. Acad Emerg Med 2021 Feb 02;28(2):184-196. [doi: 10.1111/acem.14190] [Medline: 33277724]
37. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 2004 May 01;66(3):411-421. [doi: 10.1097/01.psy.0000127692.23278.a9] [Medline: 15184705]
38. Burch VC, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and inhospital mortality. Emerg Med J 2008 Oct 01;25(10):674-678. [doi: 10.1136/emj.2007.057661] [Medline: 18843068]
39. Kim SW, Li JY, Hakendorf P, Teubner DJ, Ben-Tovim DI, Thompson CH. Predicting admission of patients by their presentation to the emergency department. Emerg Med Australas 2014 Aug 16;26(4):361-367. [doi: 10.1111/1742-6723.12252] [Medline: 24934833]
40. Kraaijvanger N, Rijpsma D, Roovers L, van Leeuwen H, Kaasjager K, van den Brand L, et al. Development and validation of an admission prediction tool for emergency departments in the Netherlands. Emerg Med J 2018 Aug 07;35(8):464-470. [doi: 10.1136/emermed-2017-206673] [Medline: 29627769]
41. Cameron A, Rodgers K, Ireland A, Jamdar R, McKay GA. A simple tool to predict admission at the time of triage. Emerg Med J 2015 Mar 13;32(3):174-179 [FREE Full text] [doi: 10.1136/emermed-2013-203200] [Medline: 24421344]

42. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018 Jul 20;13(7):e0201016 [FREE Full text] [doi: 10.1371/journal.pone.0201016] [Medline: 30028888]
43. Lucke JA, de Gelder J, Clarijs F, Heringhaus C, de Craen AJM, Fogteloo AJ, et al. Early prediction of hospital admission for emergency department patients: a comparison between patients younger or older than 70 years. Emerg Med J 2018 Jan 16;35(1):18-27. [doi: 10.1136/emermed-2016-205846] [Medline: 28814479]
44. Meisel ZF, Pollack CV, Mechem CC, Pines JM. Derivation and internal validation of a rule to predict hospital admission in prehospital patients. Prehosp Emerg Care 2008 Jul 02;12(3):314-319. [doi: 10.1080/10903120802096647] [Medline: 18584498]
45. Meisel Z, Mathew R, Wydro G, Crawford Mechem C, Pollack C, Katzer R, et al. Multicenter validation of the Philadelphia EMS admission rule (PEAR) to predict hospital admission in adult patients using out-of-hospital data. Acad Emerg Med 2009 Jun;16(6):519-525 [FREE Full text] [doi: 10.1111/j.1553-2712.2009.00422.x] [Medline: 19438413]
46. Parker CA, Liu N, Wu SX, Shen Y, Lam SSW, Ong MEH. Predicting hospital admission at the emergency department triage: a novel prediction model. Am J Emerg Med 2019 Aug;37(8):1498-1504. [doi: 10.1016/j.ajem.2018.10.060] [Medline: 30413365]
47. Sun Y, Heng B, Tay S, Seow E. Predicting hospital admissions at emergency department triage using routine administrative data. Acad Emerg Med 2011 Aug;18(8):844-850 [FREE Full text] [doi: 10.1111/j.1553-2712.2011.01125.x] [Medline: 21843220]
48. Feldman SS, Davlyatov G, Hall AG. Toward understanding the value of missing social determinants of health data in care transition planning. Appl Clin Inform 2020 Aug 26;11(4):556-563 [FREE Full text] [doi: 10.1055/s-0040-1715650] [Medline: 32851616]
49. Rubin DB. Inference and missing data. Biometrika 1976 Dec;63(3):581-592. [doi: 10.1093/biomet/63.3.581]
50. Lin W, Tsai C. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev 2019 Apr 5;53(2):1487-1509. [doi: 10.1007/s10462-019-09709-4]
51. Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005 Mar;61(1):92-105. [doi: 10.1111/j.0006-341X.2005.030814.x] [Medline: 15737082]
52. Tape TG. The area under an ROC curve. The University of Nebraska. nebraska URL: http://gim.unmc.edu/dxtests/roc3.htm [accessed 2020-10-03]
53. Wiens J, Price WN, Sjoding MW. Diagnosing bias in data-driven algorithms for healthcare. Nat Med 2020 Jan 13;26(1):25-26. [doi: 10.1038/s41591-019-0726-6] [Medline: 31932798]
54. Waljee A, Higgins P, Singal A. A primer on predictive models. Clin Transl Gastroenterol 2014 Jan 02;5(1):e44 [FREE Full text] [doi: 10.1038/ctg.2013.19] [Medline: 24384866]
55. Harrell JF. Regression Modeling Strategies With Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer; 2001:978-971.
56. Rothery P. A cautionary note on data transformation: bias in back-transformed means. Bird Study 2009 Jun 24;35(3):219-221. [doi: 10.1080/00063658809476992]
57. Pencina MJ, D'Agostino RB. Evaluating discrimination of risk prediction models: the C statistic. JAMA 2015 Sep 08;314(10):1063-1064. [doi: 10.1001/jama.2015.11082] [Medline: 26348755]
58. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. accuracy and precision of regression estimates. J Clin Epidemiol 1995 Dec;48(12):1503-1510. [doi: 10.1016/0895-4356(95)00048-8] [Medline: 8543964]
59. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996 Dec;49(12):1373-1379. [doi: 10.1016/s0895-4356(96)00236-3] [Medline: 8970487]
60. Little R, Rubin D. Statistical Analysis With Missing Data. England: John Wiley & Sons, Inc; 2002:9780471183860.
61. Vandenbroucke J, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. Int J Surg 2014 Dec;12(12):1500-1524 [FREE Full text] [doi: 10.1016/j.ijsu.2014.07.014] [Medline: 25046751]
62. Moons KM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. development, internal validation, and assessing the incremental value of a new (bio)marker. Heart 2012 May 07;98(9):683-690. [doi: 10.1136/heartjnl-2011-301246] [Medline: 22397945]
63. Steyerberg EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating. New York: Springer; 2009.
64. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. BMJ 2020 Apr 07;369:m1328 [FREE Full text] [doi: 10.1136/bmj.m1328] [Medline: 32265220]
65. Shamsoddin E. Can medical practitioners rely on prediction models for COVID-19? a systematic review. Evid Based Dent 2020 Sep;21(3):84-86 [FREE Full text] [doi: 10.1038/s41432-020-0115-5] [Medline: 32978532]

Abbreviations

AUROC: area under the receiver operating characteristics curve
CHARMS: Checklist for Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies
PROBAST: Prediction Model Risk of Bias Assessment Tool

Edited by G Eysenbach, C Lovis; submitted 28.04.21; peer-reviewed by A Vagelatos; comments to author 19.05.21; revised version received 27.05.21; accepted 28.07.21; published 16.09.21.

Please cite as:
Monahan AC, Feldman SS
Models Predicting Hospital Admission of Adult Patients Utilizing Prehospital Data: Systematic Review Using PROBAST and CHARMS
JMIR Med Inform 2021;9(9):e30022
URL: https://medinform.jmir.org/2021/9/e30022
doi:10.2196/30022
PMID:34528893

Original Paper

Predicting Health Material Accessibility: Development of Machine Learning Algorithms

Meng Ji[1], PhD; Yanmeng Liu[1], MA; Tianyong Hao[2], PhD

[1]School of Languages and Cultures, The University of Sydney, Sydney, Australia

[2]School of Computer Science, South China Normal University, Guangdong, China

Corresponding Author:
Tianyong Hao, PhD
School of Computer Science
South China Normal University
No.55 West Zhongshan Avenue, Shipai, Tianhe District
Guangdong, 510631
China
Phone: 86 15626239317
Email: haoty@m.scnu.edu.cn

Related Article:

This is a corrected version. See correction statement: https://medinform.jmir.org/2021/9/e33385

Abstract

Background: Current health information understandability research uses medical readability formulas to assess the cognitive difficulty of health education resources. This is based on an implicit assumption that medical domain knowledge represented by uncommon words or jargon form the sole barriers to health information access among the public. Our study challenged this by showing that, for readers from non-English speaking backgrounds with higher education attainment, semantic features of English health texts that underpin the knowledge structure of English health texts, rather than medical jargon, can explain the cognitive accessibility of health materials among readers with better understanding of English health terms yet limited exposure to English-based health education environments and traditions.

Objective: Our study explores multidimensional semantic features for developing machine learning algorithms to predict the perceived level of cognitive accessibility of English health materials on health risks and diseases for young adults enrolled in Australian tertiary institutes. We compared algorithms to evaluate the cognitive accessibility of health information for nonnative English speakers with advanced education levels yet limited exposure to English health education environments.

Methods: We used 113 semantic features to measure the content complexity and accessibility of original English resources. Using 1000 English health texts collected from Australian and international health organization websites rated by overseas tertiary students, we compared machine learning (decision tree, support vector machine [SVM], ensemble tree, and logistic regression) after hyperparameter optimization (grid search for the best hyperparameter combination of minimal classification errors). We applied 5-fold cross-validation on the whole data set for the model training and testing, and calculated the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy as the measurement of the model performance.

Results: We developed and compared 4 machine learning algorithms using multidimensional semantic features as predictors. The results showed that ensemble classifier (LogitBoost) outperformed in terms of AUC (0.858), sensitivity (0.787), specificity (0.813), and accuracy (0.802). Support vector machine (AUC 0.848, sensitivity 0.783, specificity 0.791, and accuracy 0.786) and decision tree (AUC 0.754, sensitivity 0.7174, specificity 0.7424, and accuracy 0.732) followed. Ensemble classifier (LogitBoost), support vector machine, and decision tree achieved statistically significant improvement over logistic regression in AUC, sensitivity, specificity, and accuracy. Support vector machine reached statistically significant improvement over decision tree in AUC and accuracy. As the best performing algorithm, ensemble classifier (LogitBoost) reached statistically significant improvement over decision tree in AUC, sensitivity, specificity, and accuracy.

Conclusions: Our study shows that cognitive accessibility of English health texts is not limited to word length and sentence length as had been conventionally measured by medical readability formulas. We compared machine learning algorithms based on semantic features to explore the cognitive accessibility of health information for nonnative English speakers. The results showed the new models reached statistically increased AUC, sensitivity, and accuracy to predict health resource accessibility for the target readership. Our study illustrated that semantic features such as cognitive ability–related semantic features, communicative actions and processes, power relationships in health care settings, and lexical familiarity and diversity of health texts are large contributors to the comprehension of health information; for readers such as international students, semantic features of health texts outweigh syntax and domain knowledge.

(JMIR Med Inform 2021;9(9):e29175) doi:10.2196/29175

KEYWORDS

health education materials; Chinese language; cognitive accessibility; readability; semantic features; health education; machine learning; prediction; accessibility; health text; cognition; accessibility; semantic; algorithm; health information

Introduction

Readability Matters

Health education materials provide important educational interventions to help increase the awareness of health risks. The recent outbreaks of the COVID-19 pandemic highlight the need to develop accessible health information, as health information appraisal has emerged as an issue in high-income countries [1]. The efficiency of health education materials largely depends on the readability and cognitive accessibility of the materials [2]. As such, the World Health Organization recommends several principles for developing health education materials regarding readability [3]. It is suggested that the readability level of medical information be lower than sixth grade for the public, and there should be easier material design for people with poor understanding capabilities [4-7]. However, studies indicate that many health education materials are more difficult than expected, leaving the layman readers encountering difficulties to comprehend the materials, which will inevitably compromise the efficiency of the health risk intervention [8-11].

Enhanced readability will improve the accessibility of health educational resources. Widely used readability assessment tools are medical readability formulas [12]. medical readability formulas measure health information readability based on word length or sentence length, assuming that the longer words and sentences are, the more difficult the health content is. These formulas are challenged by scholars due to its oversimplified factors considered in the calculations and inconsistency assessment results [13,14]. For health education texts, the cognitive difficulty in understanding medical information is caused not only by medical jargon and complex sentences but also by semantic meanings, which cannot be directly represented by word and sentence length alone [15-17]. However, readability estimation tools considering semantic features are few and underexplored. Readability estimation tools considering semantic features are in urgent need, especially for readers with better understanding of health terms yet limited exposure to English health education materials. These types of readers, represented by nonnative English speakers living in English-speaking countries, like the United States, Australia, New Zealand, or Canada, make up a large quantity of the population whose health education is of concern for the society [18-21]. These readers pose new challenges for medical readability assessment, as they normally have sufficient understanding of health terms yet limited exposure to English health education materials. In these cases, semantic features of English health texts rather than medical jargon would be suitable to estimate the cognitive accessibility of health materials.

Our study will address the challenges of using existing medical readability formulas to provide valid effective assessment of health information for readers with bilingual proficiency yet limited exposure to English health education traditions. We will introduce semantic features as indicators in cognitive accessibility evaluation. Compared with previous approaches that focus on morphological and syntactic features, we will explore the validity and effectiveness of using multidimensional semantic features (especially lexis related to English health education cultures) to analyze, model, and predict the cognitive accessibility of English health education materials. Improving cognitive accessibility of health education materials will provide a cost-effective approach to public health education. Improvement in cognitive accessibility of health education materials will contribute to social and health quality among readers from nonnative English speaking backgrounds [22].

Data Sets and Feature Extraction

Material Collection and Classification

This paper collected health education materials in English from government, health agencies, and not-for-profit organizations in Australia, considering Australia is a typical migrant country with a large amount of nonnative English speakers living in the country. The source of the health education materials includes Department of Health in state governments like Western Australia, New South Wales, and Victoria, and not-for-profit organizations [23-26]. The topic of the materials is about infectious diseases like COVID-19, Ebola, plague, or Zika, as infectious disease education is urgent in need with the background of pandemic outbreaks in recent years. In total, 1000 health education articles were collected with a size of over 500,000 words. The types of materials are patient guidelines, fact sheets, and health topics, which are health education resources accessible by the public to improve their health awareness or health knowledge. For classification, we invited 4 international students studying in Australian universities as labelers to rate the readability of the collected materials. The labelers were aged between 25 and 30 years, nonnative English speakers with advanced English skills (International English Language Testing System test score 6.5 or greater), and they were born and grew up in non-English speaking countries with limited exposure to English health education materials. They were asked to classify the collected health texts independently into easy versus hard to understand categories, and the interrater agreement was high (Cohen kappa 0.705). The final classification contained two sets of texts: easy (n=495) versus difficult (n=505; original annotated data sets in Multimedia Appendix 1).

Material Annotation and Semantic Feature Extraction

The UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System (USAS) was adopted to annotate health education materials and extract semantic features [27]. The system relies on several disambiguation methods including part-of-speech tagging, general likelihood ranking, multiword expression extraction, domain of discourse identification, and contextual rules, providing high annotation accuracy of English texts. USAS categorizes English words into 21 semantic groups, including general and abstract terms (group A); physical condition and bodily processes (group B); emotions (group D); food and drinks (group F); governmental activities (group G); residence, buildings, and habitats (group H); work and employment (group I); entertainment, sports, and activities (group K); life and living things (group L); movement, location, and transport (group M); numbers and measurements (group N); substances, materials, objects, and equipment (group O); education (group P); linguistic actions, states, and processes (group Q); social states, actions, and processes (group S); time (group T); geographical terms (group W); psychological actions, states, and processes (group X); science and technology (group Y); and names and grammatical words (group Z). With USAS, we collected 113 semantic features. In this study, we extracted these semantic features automatically from specialized English health materials to provide additional text information for developing machine learning algorithms.

Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts

Multimedia Appendix 1 shows the results of a logistic regression of the entire annotated database. A total of 26 of the 113 semantic features were identified as statistically significant features contributing to the binary classification of health texts in terms of their understandability to the target readerships, international students in tertiary education. Several semantic features contributing to the higher understandability of health texts were identified. First, *informational coherence* through pronouns (Z8) is a large contributor to the cognitive accessibility of English health texts among non-English readers, even those with advanced English language skills. The *P* and the effect size of the semantic feature Z8 were <.001 and .91, respectively, suggesting a very significant difference between easy and difficult health texts in terms of the use of pronouns. The mean score of Z8 in health texts of higher understandability was 52.84, this dropped to 20.48 in health texts of low understandability. Further, in the logistic regression analysis, the odds ratio of Z8 (ratio of odds between difficult and easy texts, with easy text as the reference text class) was 0.928 (95% CI 0.905-0.951), indicating, with the increase of 1 standard unit of Z8, the odds of the health text being a difficult health reading over the odds of the text being an easy reading was 0.928. In terms of percentage change, the odds of the health text being a difficult text was 0.031 lower than the odds of the text being an easy reading for the target readers. Semantic features related to the *logical structure* (Z7 conditional expressions such as if) were identified as statistically significant (*P*=.01). The odds ratio of Z7 was 0.86 (95% CI 0.767-0.964), indicating that holding other textual features unchanged, with the increase of one word in the Z7 class, the odds of the health text being a difficult text was 86% over the odds of the text being an easy reading.

The logistic regression result (Multimedia Appendix 1) also identified 12 semantic features as statistically significant contributors to the perceived difficulty of English health texts. Typical examples were B3 (medicines and medical treatment; odds ratio Exp(B) 1.041, 95% CI 1.012-1.071; *P*=.005), Z99 (out-of-dictionary words; odds ratio Exp(B) 1.011, 95% CI 1.004-1.018; *P*=.001), L2 (living creatures: animals, microorganism, virus, bacteria, etc; odds ratio Exp(B) 1.080, 95% CI 1.005-1.162; *P*=.036), and W5 (environmental terms: pollutants, carcinogens, inhalable particles, etc.; odds ratio Exp(B) 2.441, 95% CI 1.173-5.077; *P*=.017). These semantic features measured *lexical familiarity and diversity* of English health texts, which is another important dimension of the assessment of medical and health lexis understandability. For example, the relatively large odds ratios (2.441, 95% CI 1.173-5.077) of W5 encompassing terms related to environmental exposure and health risks indicates that, with the increase of one word in this particular category, the odds of a health text being a difficult text over the odds of the text being an easy text for the target readers was 2.441, or in terms of percentage change, this represents an increase of 144.1% of the text from an easy text to a very difficult health reading. To a lesser extent, the odds ratio of 1.080 of L2 (living creatures including microorganisms) indicates that with the increase of one word in this class, the perceived difficulty level (hard-to-understand class) of the health text increased by a mean 8.0% (95% CI 0.5%-16.2%) depending on the vocabulary range of English health terms of the readers. Semantic features relating more *abstract concepts and higher cognitive abilities* were detected as statistically significant contributors to the perceived difficulty of health texts. These include A11 (abstract terms denoting importance, significance, noticeability, or markedness; odds ratio 1.219, 95% CI 1.070-1.388; *P*=.003). This means that with the increase of one unit in the A11 class, the odds of the health text being seen as a hard-to-understand text over the text being seen as an easy text was 1.219, or an increase of 21.9%.

In the next section, we will use these predictor variables to compare the performance of machine learning algorithms in analyzing and predicting the cognitive accessibility of English health materials for the intended readership of international tertiary students.

Methods

Using machine learning algorithms and natural language processing tools to analyze and predict the understandability levels of health information has been gaining momentum. Zheng and Yu [28] used surface text features and word embeddings to support vector machine (SVM) algorithms to assess and rank the readability levels of electronic health records and Wikipedia articles. Venturi et al [29] also applied SVM to evaluate and predict the cognitive difficulty of medical informed consent forms in Italian. They used natural language features such as part of speech, type token ratio, noun verb ratio, average parse tree depth, main versus subordinate clauses distribution, distribution of verbal roots with explicit subject, and other syntactic and grammatical features related to Italian linguistic complexity. However, few existing studies have explored the effects of semantic features on the understandability of health information as our study did.

The four machine learning methods used in this study were ensemble classifier, SVM, decision tree classifier, and logistic regression classifier. Ensemble classifier (LogitBoost), SVM, and decision tree are optimizable models, as their hyperparameters can be fine-tuned through automatic grid searches to achieve minimal classification errors. For a decision tree classifier, the best-point hyperparameters (Figure 1) were the maximum number of tree splits (n=22) based on maximum deviance reduction. The observed minimal classification error of the optimized decision tree model was 0.215. For an ensemble classifier, the best-point hyperparameters (Figure 2) reached an observed minimum classification error of 0.168. The optimized hyperparameters were the ensemble method (LogitBoost), number of learners (n=210), learning rate (0.1), and maximum number of splits (n=22). For SVM, the best-point hyperparameters (Figure 3) were box constraint level (0.1), kernel function (cubic). The observed minimum classification error was 0.1944, lower than the optimized decision tree model (with a difference of 0.0206) but higher than the optimized ensemble classifier (with a difference of 0.0264).

Figure 1. Hyperparameter tuning (decision tree).

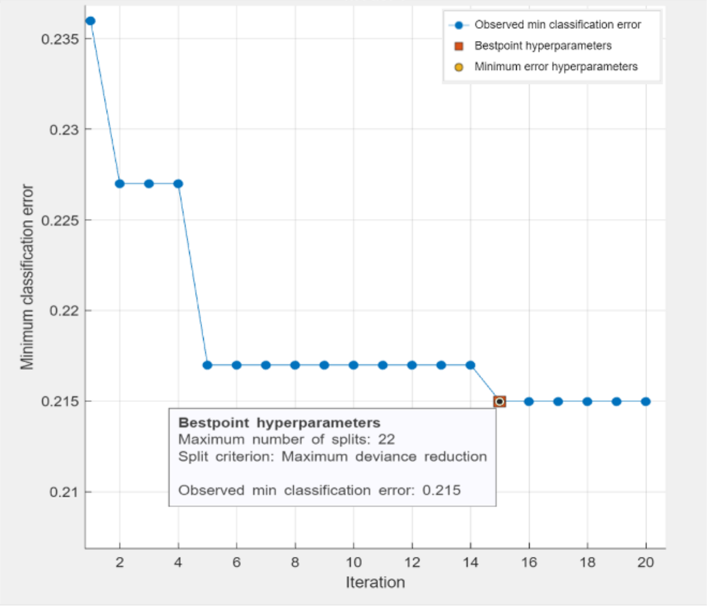


Figure 2. Hyperparameter tuning (ensemble classifier).

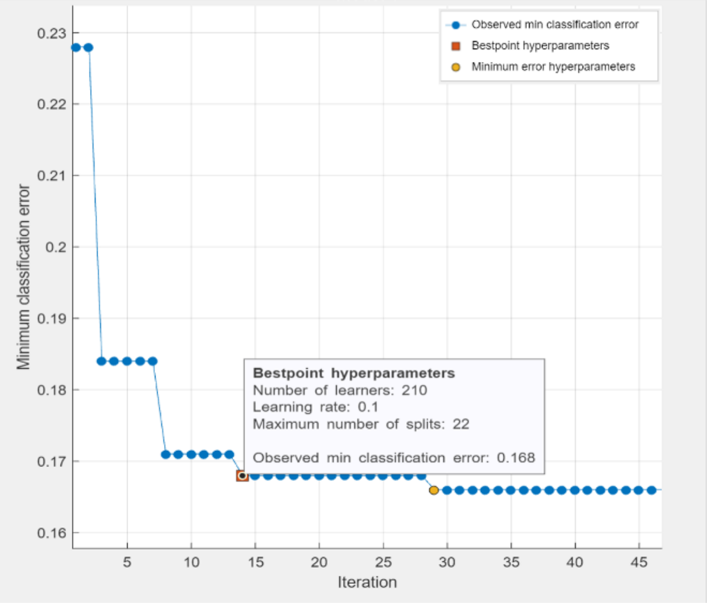
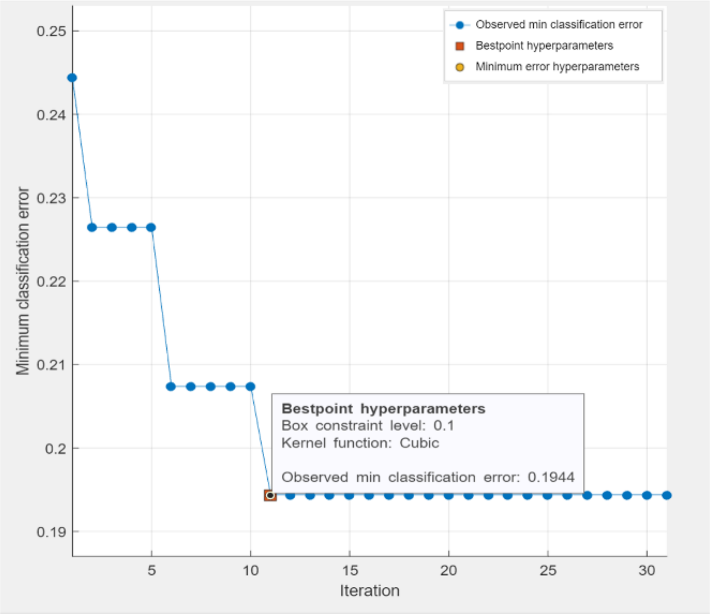


Figure 3. Hyperparameter tuning (support vector machine).



Results

The predictive performance of the four machine learning algorithms using multidimensional semantic features as predictor variables is shown in Table 1, and the results of the pairwise corrected resampled *t* test are shown in Tables 2-5. The mean scores and standard deviations of the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy were obtained through 5-fold cross-validation. The cross-validation divided the entire data set into 5 folds of equal size. In each iteration, 4 folds were used for the training data, and the remaining fold was used as the testing data. As a result, on completion of the 5-fold cross-validation, each fold was used as the testing data exactly once. We used paired-sample comparisons to investigate the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy differences of four machine learning algorithms (n=6; α=.05).

Table 1. Performance of the machine learning models using multidimensional semantic features as predictors.

Algorithm	AUC[a], mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	Accuracy, mean (SD)
LR[b]	0.614 (0.0554)	0.6282 (0.0597)	0.5724 (0.0733)	0.6010 (0.0523)
SVM[c]	0.848 (0.0172)	0.7830 (0.0368)	0.7910 (0.0420)	0.7860 (0.0153)
DT[d]	0.754 (0.0377)	0.7174 (0.0719)	0.7424 (0.0589)	0.732 (0.0317)
ENS[e]	0.858 (0.041)	0.787 (0.057)	0.813 (0.046)	0.802 (0.032)

[a] AUC: area under the operating characteristic curve.

[b] LR: logistic regression.

[c] SVM: support vector machine.

[d] DT: decision tree.

[e] ENS: ensemble classifier (LogitBoost).

Table 2. Pairwise corrected resampled *t* test of area under the curve differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.2340 (0.0669)	0.0299	–0.3171 to –0.1509	–7.817 (4)	.001
LR vs DT[c]	–0.1460 (0.0551)	0.0246	–0.2144 to –0.0777	–5.931 (4)	.004
LR vs ENS[d]	–0.2440 (0.0564)	0.0252	–0.3140 to –0.1740	–9.675 (4)	.001
SVM vs DT	0.0880 (0.0192)	0.0086	–0.0641 to 0.1119	10.230 (4)	.001
SVM vs ENS	–0.0100 (0.0374)	0.0167	–0.0565 to –0.0365	–0.598 (4)	.582
DT vs ENS	–0.0980 (0.0192)	0.0086	–0.1219 to –0.0741	–11.392 (4)	<.001

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 3. Pairwise corrected resampled *t* test of sensitivity differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.1548 (0.0303)	0.0135	–0.1924 to –0.1172	–11.429 (4)	<.001
LR vs DT[c]	–0.1002 (0.0720)	0.0322	–0.1896 to –0.0108	–3.111 (4)	.036
LR vs ENS[d]	–0.1588 (0.0945)	0.0423	–0.2761 to –0.0414	–3.756 (4)	.020
SVM vs DT	0.0546 (0.0697)	0.0312	–0.0319 to 0.1411	1.752 (4)	.155
SVM vs ENS	–0.0040 (0.0855)	0.0382	–0.1102 to –0.1022	–0.105 (4)	.922
DT vs ENS	–0.0586 (0.0371)	0.0166	–0.1046 to –0.0126	–3.535 (4)	.024

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 4. Pairwise corrected resampled *t* test of specificity differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.2186 (0.0968)	0.0433	–0.3389 to –0.0984	–5.047 (4)	.007
LR vs DT[c]	–0.1720 (0.0822)	0.0368	–0.2741 to –0.0699	–4.679 (4)	.009
LR vs ENS[d]	–0.2410 (0.0677)	0.0303	–0.3251 to –0.1569	–7.959 (4)	.001
SVM vs DT	0.0466 (0.1059)	0.0474	–0.0849 to 0.1781	0.984 (4)	.381
SVM vs ENS	–0.0224 (0.0918)	0.0411	–0.1364 to –0.0916	–0.545 (4)	.614
DT vs ENS	–0.0690 (0.0334)	0.0149	–0.1105 to –0.0275	–4.619 (4)	.010

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 5. Pairwise corrected resampled *t* test of accuracy differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.1850 (0.0507)	0.0227	–0.2480 to –0.1220	–8.152 (4)	.001
LR vs DT[c]	–0.1370 (0.0482)	0.0215	–0.1968 to –0.0771	–6.360 (4)	.003
LR vs ENS[d]	–0.2010 (0.0549)	0.0246	–0.2692 to –0.1328	–8.182 (4)	.001
SVM vs DT	0.0480 (0.0295)	0.0132	0.0114 to 0.0846	3.639 (4)	.022
SVM vs ENS	–0.0160 (0.0366)	0.0164	–0.0615 to 0.0295	–0.976 (4)	.384
DT vs ENS	–0.0640 (0.0148)	0.0066	–0.0823 to –0.0457	–9.704 (4)	.001

[a] LR: logistic regression.

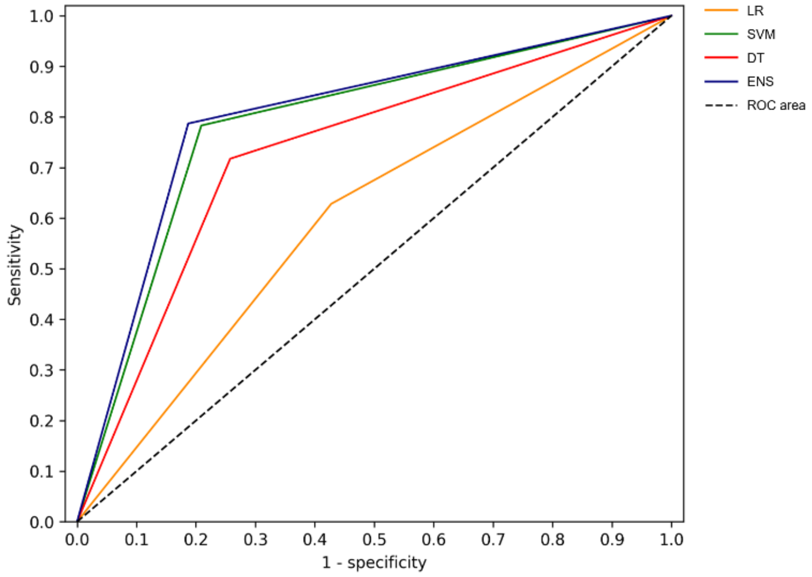
[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 2 shows that, in terms of AUC, ensemble classifier (LogitBoost), decision tree, and SVM reached statistically improved AUC over logistic regression (0.614): ensemble classifier (0.858; *P*=.001), decision tree (0.754; *P*=.004), and SVM (0.848, *P*=.001). In terms of sensitivity (Table 3), ensemble classifier (0.787, *P*=.020), decision tree (0.7174, *P*=.036), and SVM (0.783; *P*<.001) reached statistically significant improvement over logistic regression (0.6282). In terms of model specificity (Table 4), ensemble classifier, decision tree, and SVM all reached statistically improved specificity over logistic regression (0.5724): ensemble classifier (0.813; *P*=.001), decision tree (0.7424; *P*=.009), and SVM (0.791; *P*=.007). Lastly, with regard to model overall accuracy (Table 5), again, LogitBoost, decision tree, and SVM outperformed logistic regression (0.601): ensemble classifier (0.802; *P*=.001), decision tree (0.732; *P*=.003), and SVM (0.786; *P*=.001). Comparing SVM, ensemble classifier and decision tree, the former two algorithms outperformed decision tree consistently in AUC (*P*=.001 and *P*<.001, respectively), and accuracy (*P*=.022 and *P*=.001, respectively). Only ensemble classifier outperformed decision tree significantly in terms of model sensitivity (*P*=.024), and specificity (*P*=.010), using the paired-sample comparisons (n=6; α=.05). These results suggest that, when using semantic features as predictor variables, the most stable and highest-performing algorithm is ensemble classifier (LogitBoost), followed by SVM. Ensemble classifier, decision tree, and SVM all achieved statistically significant improvement over logistic regression in AUC, specificity, sensitivity, and accuracy. SVM did not improve significantly over decision tree in terms of sensitivity and specificity, but ensemble classifier did. Overall, the best AUC, sensitivity, specificity, and accuracy were achieved by LogitBoost as an ensemble classifier (Figure 4).

Figure 4. Mean ROC curve for machine learning algorithms. DT: decision tree; LR: logistic regression; ROC: receiver operating characteristic; SVM: support vector machine.



Discussion

Principal Findings

The understandability of health texts has long been assessed using medical readability formulas. This has simplified and limited the discussion of health information accessibility to two known barriers (ie, medical jargon and syntactic features). Existing research has been limited in exploring these issues despite methodological innovation in applying and leveraging machine learning algorithms and natural language processing tools in this field. Our study explored health information accessibility using semantic features of health information that are less studied. This was in line with clinical insights into patient-oriented health education, which identified multiple textual features as highly relevant to the understanding of specialized health information. However, few existing studies have attempted to translate recent clinical guidelines and insights to quantitative computational studies using linguistic features related to the semantic content as exemplified in our study. Using semantic annotation tools, we explored effects of various semantic features on the understandability of health texts for the target readers.

In the multiple machine learning algorithm comparison, the importance of semantic features was verified. It was found that, in the algorithm comparison experiments, using multidimensional semantic features as predictor variables, LogitBoost achieved the highest performance in terms of AUC, sensitivity, specificity, and accuracy, which were statistically significant large improvements (measured in pairwise resampled *t* tests). Among the 4 algorithms used, AUCs, sensitivity, specificity, and accuracy were consistently high when using multidimensional semantic features as predictors variables. This finding suggests that multidimensional semantic features are large contributors to the cognitive accessibility of English health texts among readers with English proficiency but limited exposure to English health education traditions (indicated by less familiarity of relevant health lexis and abstract concepts).

Considering that the readership under study were educated international tertiary students who had less barriers to understand and analyze complex English syntactic structures but had limited exposure to English-based health education environments, our study shows that, for readers from this background of health literacy and education level, informational coherence and logical structure were large contributors to the ease of health texts. Features of health-related lexical familiarity and diversity or those indicating abstract concepts or requiring higher cognitive abilities can significantly increase the difficulty of English health information for readers from non-English speaking and distinct health education backgrounds, despite their English proficiency from tertiary education.

In the development of effective reader-oriented health educational resources, enhancing semantic features, which were identified as large contributors to cognitive ease, can lead to more beneficial reading experiences among the target readers. Textual interventions can be effectively introduced to reduce the cognitive load of health texts, such as health lexical diversity (especially those of large odds ratios such as environmental exposure and health risks), or those requiring higher cognitive abilities, such as abstract terms denoting importance, significance, noticeability or markedness of health events and situations. These semantic features can significantly increase the difficulty and inaccessibility of English health education resources among international students, as these semantic features require greater, more sustained exposure to English public health education traditions.

Limitations and Future Research

Our study was based on a small group of international students from native Chinese speaking backgrounds. Their rating of the cognitive understandability of English health texts could have been biased by their shared cultural backgrounds. This was, however, intended to control for cultural demographic diversity in our study. Whether this finding applies to other cohorts of international tertiary students remains to be evaluated through similar experiment design. Another considerable limitation of our study is the lack of explanation by the machine learning–based prediction. In future research, we aim to develop more explainable machine learning models to increase the interpretability of the prediction results.

Conclusion

Our study showed that cognitive accessibility of English health texts is not limited to medical jargon and complex syntax such as long words and sentences conventionally measured by medical readability formulas. We compared machine learning algorithms using multiple semantic features to explore the cognitive accessibility of health information from multiple semantic perspectives. The results showed the strength of our models in terms of consistently high AUC, sensitivity, specificity, and accuracy to predict health resource accessibility for the target readers, indicating that semantics contribute to the comprehension of health information and that, for readers with advanced education, semantic features that underpin the English-based health education can outweigh syntax and specialized medical domain knowledge.

Authors' Contributions

MJ and TH were responsible for overall research design; MJ was responsible for paper writing and revision, and YL was responsible for formal analysis and data curation.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Variables in the logistic regression of health text understandability membership.
[DOCX File , 34 KB - medinform_v9i9e29175_app1.docx]

References

1. Farooq A, Laato S, Islam AKMN. Impact of online information on self-isolation intention during the COVID-19 pandemic: cross-sectional study. J Med Internet Res 2020 May 06;22(5):e19128 [FREE Full text] [doi: 10.2196/19128] [Medline: 32330115]
2. Gal I, Prigat A. Why organizations continue to create patient information leaflets with readability and usability problems: an exploratory study. Health Educ Res 2005 Aug;20(4):485-493. [doi: 10.1093/her/cyh009] [Medline: 15613490]
3. WHO Strategic Communications Framework for effective communications. World Health Organization. 2017 Mar 30. URL: https://www.who.int/mediacentre/communication-framework.pdf [accessed 2021-04-20]
4. Weiss B. Health literacy. American Medical Association. 2003. URL: http://lib.ncfh.org/pdfs/6617.pdf [accessed 2021-04-20]
5. Weiss BD, Coyne C. Communicating with patients who cannot read. N Engl J Med 1997 Jul 24;337(4):272-274. [doi: 10.1056/NEJM199707243370411] [Medline: 9227936]
6. Cotugna N, Vickery CE, Carpenter-Haefele KM. Evaluation of literacy level of patient education pages in health-related journals. J Community Health 2005 Jun;30(3):213-219. [doi: 10.1007/s10900-004-1959-x] [Medline: 15847246]
7. Doak L, Doak C, Meade C. Strategies to improve cancer education materials. Oncol Nurs Forum 1996 Sep;23(8):1305-1312. [Medline: 8883075]
8. Smale M, Renfrew MJ, Marshall JL, Spiby H. Turning policy into practice: more difficult than it seems. The case of breastfeeding education. Matern Child Nutr 2006 Apr;2(2):103-113 [FREE Full text] [doi: 10.1111/j.1740-8709.2006.00045.x] [Medline: 16881920]
9. Horner SD, Surratt D, Juliusson S. Improving readability of patient education materials. J Community Health Nurs 2000;17(1):15-23. [doi: 10.1207/S15327655JCHN1701_02] [Medline: 10778026]
10. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res 2010 Oct;468(10):2572-2580 [FREE Full text] [doi: 10.1007/s11999-010-1380-y] [Medline: 20496023]
11. Williams AM, Muir KW, Rosdahl JA. Readability of patient education materials in ophthalmology: a single-institution study and systematic review. BMC Ophthalmol 2016 Aug 03;16:133 [FREE Full text] [doi: 10.1186/s12886-016-0315-0] [Medline: 27487960]
12. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996 Feb;1(1):7-28. [doi: 10.1080/13548509608400003]
13. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. AMIA Annu Symp Proc 2007 Oct 11:418-422 [FREE Full text] [Medline: 18693870]
14. Pichert JW, Elam P. Readability formulas may mislead you. Patient Education Counseling 1985 Jun;7(2):181-191. [doi: 10.1016/0738-3991(85)90008-4]
15. Sand-Jecklin K. The impact of medical terminology on readability of patient education materials. J Community Health Nurs 2007;24(2):119-129. [doi: 10.1080/07370010701316254] [Medline: 17563283]
16. Rosemblat G, Logan R, Tse T, Graham L. Text features and readability: expert evaluation of consumer health text. 2006 Presented at: Mednet 2006: 11th World Congress on Internet in Medicine the Society for Internet in Medicine; October 14-16, 2006; Toronto, Canada.
17. Masoni M, Guelfi MR. Going beyond the concept of readability to improve comprehension of patient education materials. Intern Emerg Med 2017 Jun;12(4):531-533. [doi: 10.1007/s11739-017-1645-5] [Medline: 28260222]
18. Mladovsky P. Migrant health in the EU. Eurohealth. 2007. URL: https://www.researchgate.net/publication/279187422_Migrant_Health_in_the_EU [accessed 2021-04-20]
19. Gondek M, Shogan M, Saad-Harfouche FG, Rodriguez EM, Erwin DO, Griswold K, et al. Engaging immigrant and refugee women in breast health education. J Cancer Educ 2015 Sep;30(3):593-598 [FREE Full text] [doi: 10.1007/s13187-014-0751-6] [Medline: 25385693]
20. Frost E, Markham C, Springer A. Refugee health education: evaluating a community-based approach to empowering refugee women in Houston, Texas. Adv in Soc Work 2018 Sep 18;18(3):949-964. [doi: 10.18060/21622]
21. Rosenthal DA, Russell J, Thomson G. The health and wellbeing of international students at an Australian university. High Educ 2006 Oct 5;55(1):51-67. [doi: 10.1007/s10734-006-9037-1]
22. Report on the health of refugees and migrants in the WHO European Region. World Health Organisation. 2018. URL: https://www.euro.who.int/en/publications/abstracts/report-on-the-health-of-refugees-and-migrants-in-the-who-european-region-no-public-health-without-refugee-and-migrant-health-2018 [accessed 2021-04-20]
23. Health on the Net. URL: https://www.hon.ch/en/ [accessed 2021-04-20]
24. Australian Government Department of Health. URL: https://www.health.gov.au [accessed 2021-03-25]
25. NSW Health. URL: https://www.health.nsw.gov.au [accessed 2021-03-25]
26. Department of Health, Government of Western Australia. URL: https://ww2.health.wa.gov.au [accessed 2021-03-25]

27. Rayson P, Archer D, Piao S, McEnery A. The UCREL semantic analysis system. eprints.lancs.ac.uk. 2004 Jun. URL: https://www.researchgate.net/publication/228881331_The_UCREL_semantic_analysis_system [accessed 2021-04-20]
28. Zheng J, Yu H. Assessing the readability of medical documents: a ranking approach. JMIR Med Inform 2018 Mar 23;6(1):e17 [FREE Full text] [doi: 10.2196/medinform.8611] [Medline: 29572199]
29. Venturi G, Bellandi T, Dell'Orletta F, Montemagni S. NLP–based readability assessment of health–related texts: a case study on Italian informed consent forms. In: Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis. 2015 Presented at: Sixth International Workshop on Health Text Mining and Information Analysis; 2015; Lisbon, Portugal p. 131-141. [doi: 10.18653/v1/w15-2618]

Abbreviations

AUC: area under the operating characteristic curve
SVM: support vector machine
UCREL: University Centre for Computer Corpus Research on Language
USAS: UCREL Semantic Analysis System

Edited by C Lovis; submitted 29.03.21; peer-reviewed by D Carvalho; comments to author 26.04.21; revised version received 26.04.21; accepted 25.07.21; published 01.09.21.

Please cite as:
Ji M, Liu Y, Hao T
Predicting Health Material Accessibility: Development of Machine Learning Algorithms
JMIR Med Inform 2021;9(9):e29175
URL: https://medinform.jmir.org/2021/9/e29175
doi:10.2196/29175
PMID:34468321

Review

Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review

Wael Abdelkader[1*], MD, MSc; Tamara Navarro[1*], MLiS; Rick Parrish[1*], DiplT; Chris Cotoi[1*], BEng, EMBA; Federico Germini[1,2*], MD, MSc; Alfonso Iorio[1,2*], MD, PhD, FRCPC; R Brian Haynes[1,2*], MD, PhD; Cynthia Lokker[1*], MSc, PhD

[1]Health Information Research Unit, Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
[2]Department of Medicine, McMaster University, Hamilton, ON, Canada
[*]all authors contributed equally

Corresponding Author:
Wael Abdelkader, MD, MSc
Health Information Research Unit
Department of Health Research Methods, Evidence, and Impact
McMaster University
1280 Main St W
CRL Building, First Floor
Hamilton, ON, L8S 4K1
Canada
Phone: 1 647 563 5732
Email: Abdelkaw@mcmaster.ca

Abstract

Background: The rapid growth of the biomedical literature makes identifying strong evidence a time-consuming task. Applying machine learning to the process could be a viable solution that limits effort while maintaining accuracy.

Objective: The goal of the research was to summarize the nature and comparative performance of machine learning approaches that have been applied to retrieve high-quality evidence for clinical consideration from the biomedical literature.

Methods: We conducted a systematic review of studies that applied machine learning techniques to identify high-quality clinical articles in the biomedical literature. Multiple databases were searched to July 2020. Extracted data focused on the applied machine learning model, steps in the development of the models, and model performance.

Results: From 3918 retrieved studies, 10 met our inclusion criteria. All followed a supervised machine learning approach and applied, from a limited range of options, a high-quality standard for the training of their model. The results show that machine learning can achieve a sensitivity of 95% while maintaining a high precision of 86%.

Conclusions: Machine learning approaches perform well in retrieving high-quality clinical studies. Performance may improve by applying more sophisticated approaches such as active learning and unsupervised machine learning approaches.

(JMIR Med Inform 2021;9(9):e30401) doi:10.2196/30401

KEYWORDS

machine learning; bioinformatics; information retrieval; evidence-based medicine; literature databases; systematic review; accuracy; medical literature; clinical support; clinical care

Introduction

Background and Significance

Evidence-based medicine (EBM) is identified by three key elements: the best available clinical evidence, clinician expertise, and application of the evidence with consideration of patients' circumstances, values, and preferences [1]. EBM complements or reduces reliance on expert opinion with a coherent and structured framework for assessing and applying the best evidence to patient care decisions [2]. An obvious and worsening barrier to the implementation of EBM is the continuously growing body of medical literature. According to the National

Library of Medicine, over 900,000 new citations were indexed in MEDLINE in 2020, very few of which were relevant to or ready for clinical attention [3]. Searching for the best clinical care evidence is a challenging task for researchers and clinicians, and facilitation of the search process is a necessity [4].

Search Filters

Search filters, also referred to as hedges, allow researchers, clinicians, and librarians to retrieve evidence from bibliographic databases and journals by filtering searches to return reliable and specific articles to address clinical questions, produce systematic reviews, or inform clinical guidelines [5]. MEDLINE search filters, for example, enable researchers to combine the use of free text with controlled vocabularies like Medical Subject Heading (MeSH) terms and other indexing features to improve search results targeting the clinical question at hand [6,7]. There are search filters that focus on the purpose of a study and its methods or topical content areas [8]. Topical search filters help identify articles based on particular clinical conditions using terms related to that condition [8], while methodological search filters comprise terms that identify articles based on their research purpose [9]. For example, the Hedges project, developed by the Health Information Research Unit at McMaster University, provides search filters for MEDLINE, PsycINFO, and EMBASE using the OVID syntax for a range of purpose categories of articles such as treatment, diagnosis, and prognosis and include methodological terms [4,10,11]. For searches seeking articles on a treatment (purpose), the search hedge includes methodological terms related to clinical or randomized controlled trials (RCTs), while the diagnosis search hedge includes methodological terms including sensitivity and specificity [12].

These search filters were developed to identify high-quality studies based on established critical appraisal criteria for methodological rigor [13-15]. This was done by annotating articles as meeting or not meeting criteria and using the annotated dataset to evaluate the performance of search terms to optimally retrieve the high-quality studies. For RCTs, applying the Cochrane risk for bias tool includes assessing randomization method, allocation concealment, follow-up data for at least 80% of participants, blinding of participants, and outcome assessors [14]. For the Hedges project, the criteria applied to articles by purpose are available online [15].

Clinical search filters are intended to help clinicians, researchers, and policymakers quickly access relevant studies and systematic reviews in a way that can be tailored to the user's demand [8]. The filters differ in their sensitivity and specificity according to the terms used, databases searched, and precision of the filter [16]. Some filters offer high specificity, which limits the proportion of off-target articles that are retrieved. This is useful for busy clinicians who value the most efficient use of their time in finding relevant evidence quickly. Search filters may also have the option to maximize sensitivity and identify all potentially relevant articles at the cost of including a higher proportion of off-target articles [17], an approach more suited to the conduct of systematic literature reviews.

Although search filters, such as Clinical Queries in PubMed, have been used since 1990 and have continued to work well over the years [18], they have some limitations. One limitation is their partial dependence on MeSH indexing terms, as the process of indexing of articles within MEDLINE can take up to a year for some articles [19]. For diagnostic studies, there is large variability in designs and methods, which may result in largely incomplete literature searches [7]. When applied in the context of conducting a systematic review, the highly specific filters result in missing evidence [7], and the high sensitivity search filters will only partially reduce the time-consuming task of screening retrieved titles and abstracts [20].

Overview of Machine Learning Applied for Text Processing

Machine learning is a subset of artificial intelligence that refers to a series of computational methods using experience to improve performance or achieve accurate and precise predictions. Experience, in this context, refers to the information made available to the machine for the analysis [21]. A more detailed definition was provided by Mitchell [22]: "A computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T, as measured by P, improves with experience E."

Machine learning applications have become increasingly popular and essential in health care [23], as the system generates an enormous amount of data every day [24]. Machine learning can identify relevant relations in large health care–generated datasets and derive algorithms that generate accurate predictions [25,26]. For example, machine learning has been used to predict the risk for nosocomial infection by leveraging data from electronic health records [27-29]. A machine learning classifier is a mathematical procedure responsible for identifying the patterns and performing the prediction task on the dataset, while a machine learning model is the output of the algorithm [30]. A machine learning model represents the complete learning process including the training of the algorithm and the used set of features [30].

Another application of machine learning in the health care and biomedical literature is text mining, which refers to the discovery of previously unknown information from unstructured textual data [31]. This is done by converting the text to structured analyzable data using natural language processing (NLP) [32]. With the exponential increase in the amount of information available for clinicians and researchers, both in biomedical literature and electronic health records [33], text mining has been applied for text summarization [34], literature retrieval [35], and evidence grading [36]. Machine learning has also been applied to automate the screening process for systematic reviews, identifying relevant articles while decreasing workload and increasing efficiency [20,37,38]. Semantic analysis, the process of understanding text by interpreting meanings from the unstructured text [39], has been applied to information extraction from the biomedical literature [40].

There are several types of machine learning determined by their mathematical approach [41]. The basic machine learning strategies are supervised learning, unsupervised learning, and reinforced learning [41,42]. Supervised learning relies on a prelabeled training dataset to provide the machine with the necessary input to make accurate predictions [41]. Decision tree (DT), naïve Bayes (NB), and support vector machine (SVM) are common supervised machine learning algorithms [43]. Unsupervised learning does not use labeled data and is mainly used for structuring and organizing data rather than classification [43]. In reinforced learning, the algorithm learns by reacting to its environment and reaches predictions via a reward system [42]. A common machine learning technique is ensemble learning, which combines more than one classifier to perform an individual prediction task. Boosting is one of the commonly used ensemble learners, which combines multiple weak classifiers and converts them into one strong classifier [41]. Neural networks are multilayer mathematical structures consisting of an input layer, an output layer, and a hidden layer (commonly more than one layer) in between [44]. In each layer a series of calculations occurs, leading to better performance [44]. Due to the multilayer nature of neural networks, their field of study is known as deep learning. Neural networks can be supervised, unsupervised, or reinforced [45].

Another appealing application of machine learning approaches to the biomedical literature is to improve retrieval of clinically relevant articles, building on and hopefully overcoming the limitation faced by Boolean searching. Several studies have been conducted to assess the performance of machine learning classifiers to identify specific categories of published articles. For example, Marshall and colleagues [46] applied machine learning to identify RCTs. Del Fiol and colleagues [35] used machine learning to extract only scientifically sound treatment studies from PubMed. However, no systematic review of studies objectively assessing the performance of such machine learning models, ideally comparing their performance to traditional evidence retrieval methods such as validated Boolean search filters or manual critical appraisal by experts in the field, has been performed to date. Such a systematic review would be of critical value in driving future machine learning research aimed at improving the delivery of relevant evidence to the point of care.

Objective

The objective of this systematic review is to summarize the nature (methods and approaches) and comparative performance (eg, recall and precision) of machine learning approaches that have been applied to retrieve high-quality evidence for clinical consideration from the biomedical literature. High-quality is defined as articles that meet established methodological critical appraisal criteria, with annotated datasets that apply these criteria considered the gold standard.

Methods

The following subsections describe in detail the steps that were conducted to identify, screen, and abstract data from the included studies.

Search Strategies

Nine databases were searched from inception to July 8, 2020, to identify relevant articles: Web of Science (title, abstract); MEDLINE; Embase; PsychINFO (title, abstract, keyword, subject terms); Wiley Online Library; ScienceDirect (title, abstract, keyword); CINAHL; IEEE (title, abstract, keywords), and Association of Computer Machinery digital library (title, abstract). The Multidisciplinary Digital Publishing Institute (title, abstract) database was searched on November 17, 2020. The search strategy was developed with a librarian (TN). Search terms related to 4 concepts—machine learning, literature retrieval, high research quality, and biomedical literature—were combined using the AND Boolean operator. The OVID MEDLINE search included the following terms, which were translated for the other databases (mp = multipurpose, searching within the title, original title, abstract, subject heading, name of substance, and registry word fields):

- Machine learning: (neural networks/ or machine learning/ or natural language processing/ or data mining/ or support vector machine/ or ("text categorization" or "text classification" or "text analysis" or "literature mining" or "text mining").mp)
- Study objective or goal: ("Abstracting and Indexing"/ or "information storage and retrieval"/ or ("article retrieval" or "literature surveillance" or "literature screening" or "article screening" or "evidence search" or "evidence screening" or "evidence review" or "information retrieval" or "literature survey" or "document classification" or "review efficiency" or "citation screening" or "literature databases").mp.)
- High-quality: ("Sensitivity and Specificity"/ or evidence-based medicine/ or ("quality" or "evidence" or "high-quality" or "clinical trial" or "random*" or "randomized controlled trial" or "sensitivity or specificity" or "accuracy" or "precision").mp.)

In the Association of Computer Machinery digital library and Multidisciplinary Digital Publishing Institute search queries, terms related to the biomedical literature were included: ("PubMed" or "MEDLINE" or "medical literature" or "Biomedical literature").

Study Selection

Articles retrieved by our search queries were collected in a single Research Information Systems file using JabRef software. Deduplication was conducted using both JabRef automatic deduping and Covidence automatic deduplication. We included articles that met the following criteria:

- Reported on the use of a machine learning approach for the retrieval of single studies or systematic reviews concerning the management of health care problems in large biomedical bibliographic databases such as MEDLINE and EMBASE
- Classified retrieved articles based on quality (using a gold standard)
- Used a textual analysis machine learning approach
- Evaluated the performance of the machine learning approach (ie, they present a comparison of retrieval methods or other ways of appraising the performance of the machine learning approach)
- Conducted within the biomedical literature domain
- Published in the English language

Abstract and Full-Text Screening

Titles and abstracts of all the retrieved articles were screened independently in Covidence.org by two members of the study team. Articles were assessed as relevant, irrelevant, or maybe relevant. The full texts of relevant and maybe relevant articles were then reviewed in duplicate, with conflicts adjudicated by a third team member.

Data Extraction

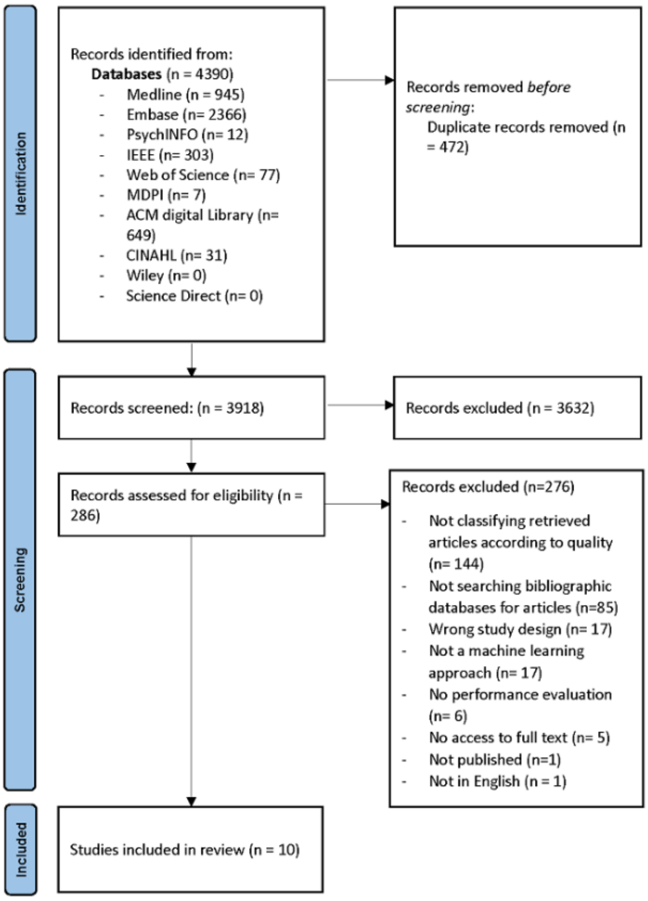
A data extraction spreadsheet was developed to gather data regarding the methods of the machine learning approaches as detailed by the survey by Agarwal and Mittal [47] and included details on preprocessing steps, text representation, feature selection, feature extraction, and classifiers used. Additionally, we extracted data specific to the retrieval of high-quality articles such as the quality gold standard, the comparators used to test the machine learning models, and the performance of the developed algorithms.

Results

Study Selection

Our search queries retrieved 3918 articles after 472 duplicates were removed; 3632 were excluded during the title and abstract screening for not applying a machine learning approach to biomedical articles. A total of 286 were selected for full-text screening, and 10 articles met our eligibility criteria (Figure 1) [48]. Due to the heterogeneity in the population (retrieved articles), index method (machine learning algorithm used), gold standard, and outcomes (definition of high-quality study), we did not perform a quantitative synthesis of the results.

Figure 1. PRISMA flow diagram of the studies identification process for the systematic review [48].


Identification
Records identified from:
Databases (n = 4390)
- Medline (n = 945)
- Embase (n= 2366)
- PsychINFO (n= 12)
- IEEE (n= 303)
- Web of Science (n= 77)
- MDPI (n= 7)
- ACM digital Library (n= 649)
- CINAHL (n= 31)
- Wiley (n= 0)
- Science Direct (n= 0)
Records removed *before screening*:
Duplicate records removed (n = 472)
Screening
Records screened: (n = 3918)
Records excluded (n = 3632)
Records assessed for eligibility (n = 286)
Records excluded (n=276)
- Not classifying retrieved articles according to quality (n= 144)
- Not searching bibliographic databases for articles (n=85)
- Wrong study design (n= 17)
- Not a machine learning approach (n= 17)
- No performance evaluation (n= 6)
- No access to full text (n= 5)
- Not published (n=1)
- Not in English (n = 1)
Included
Studies included in review (n = 10)

Quality Gold Standard

Each study used a quality gold standard database of original studies or systematic reviews that were manually reviewed and annotated by experts based on their scientific soundness and clinical relevance (Table 1). Datasets of articles that met or did not meet standards for quality and relevance were used to train the machine learning models. Four studies used the American College of Physicians (ACP) Journal Club as their quality gold standard [49-52], 3 studies used the Clinical Hedges dataset [4,35,36,53], 2 studies considered articles that were included in treatment clinical guidelines as high quality [54,55], and 1 article used the Cochrane Library as their gold standard [56].

Table 1. The quality standard used as the training dataset for developing the classifiers in the included studies.

Author	Quality gold standard
Aphinyanaphongs et al [49]	ACP[a] Journal Club (treatment class)[b]
Aphinyanaphongs et al [50]	ACP Journal Club (treatment, diagnosis, etiology, prognosis)[b]
Aphinyanaphongs et al [51]	ACP Journal Club (treatment, diagnosis, etiology, prognosis)[b]
Kilicoglu et al [53]	Clinical Hedges[b]
Lin et al [52]	ACP Journal Club (unspecified classes of articles)[b]
Afzal et al [36]	Clinical Hedges[b]
Bian et al [54]	Articles cited in 11 clinical guidelines on the treatment of cardiac, autoimmune, and respiratory diseases
Del Fiol et al [35]	Clinical Hedges[b]
Bian et al [55]	Articles cited in 11 clinical guidelines on the treatment of cardiac, autoimmune, and respiratory diseases
Afzal et al [56]	Cochrane Library Reviews

[a] ACP: American College of Physicians.

[b] Hand searches of articles from approximately 125 clinical journals that were assessed by critical appraisal criteria; articles meeting criteria were then judged by clinicians for clinical relevance. ACP Journal Club includes additional reviews by clinicians.

Preprocessing Methods

A matrix of the preprocessing steps that were applied to the dataset before developing the classifiers as reported in the included studies is presented in Table 2. Seven of the included studies provided details of their preprocessing steps [35,36,49-51,53,56], which included the conversion of text to lowercase, word-stemming, and removal of stop words. Additionally, 6 studies applied a term weighting method [36,49-51,53,56] to express the importance of a word in each document based on its frequency. Afzal et al [36] used vocabulary pruning by removing off topic-specific frequent terms and rarely occurring terms. Three studies did not specify the steps for their preprocessing steps [52,54,55].

Table 2. Preprocessing steps applied to article data for preparing the datasets for machine learning algorithm development.

Author	Text converted to lowercase	Removal of punctuation	Removal of stop words	Porter- stemming	Weighting method	Unique preprocessing considered
Aphinyanaphongs et al [49]	✓[a]	✓	✓	✓	Log frequency with redundancy	NR[b]
Aphinyanaphongs et al [50]	✓	✓	✓	✓	Log frequency with redundancy	NR
Aphinyanaphongs et al [51]	✓	✓	✓	✓	Log frequency with redundancy	Removed infrequent words
Kilicoglu et al [53]	✓	NR	✓	✓	Information gain measure	Removed infrequent words
Lin et al [52]	NR	NR	NR	NR	NR	NR
Afzal et al [36]	✓	NR	✓	✓	TF-IDF[c]	Vocabulary pruning
Bian et al [54]	NR	NR	NR	NR	NR	NR
Del Fiol et al [35]	✓	NR	✓	NR	NR	Removed articles without abstracts, concatenated title, and abstract words
Bian et al [55]	NR	NR	NR	NR	NR	NR
Afzal et al [56]	✓	NR	NR	NR	TF-IDF	Removed articles with missing values

[a] Applied.

[b] NR: not reported.

[c] TF-IDF: term frequency–inverse document frequency.

Feature Selection

Most of the included articles relied on the text as their features (Multimedia Appendix 1). Seven articles used words from titles and abstracts as their features [35,36,49-51,53,56]. Kilicoglu et al [53] and Afzal et al [36] used article metadata features, Unified Medical Language System features, SemRep semantic prediction, and MeSH terms in combination with the words of titles and abstracts features. Lin et al [52] selected specific features from the citation dataset: journal impact factor, MeSH terms, sample size, *P* value, and confidence intervals. Bian et al [54,55] relied on MEDLINE metadata as well as bibliometric features, which included citation count, journal impact factor, number of comments on PubMed, Altmetric score, study sample size, registration in ClinicalTrials.gov, and article age, and assessed how each feature contributed to the classification. The experiment by Bian et al [55] used only time-agnostic features (features available at the time of an article's publication), which are journal impact factor, sample size, number of grants, number of authors, number of clinically useful sentences, scientific impact of authors' institution, numbers of references, page count, registration in ClinicalTrials.gov, and publication in PubMed Central. Afzal et al [56] used automatic feature engineering with RapidMiner software for the title and abstract text feature extraction as part of the multilayer perceptron model.

Machine Learning Classifier

The majority of the included studies developed multiple algorithms and selected the top-performing one for their main classification tasks (Table 3). Aphinyanaphongs et al [49,50], initially reported their results using SVM, NB, and boosting algorithms in both their 2003 and 2005 experiments; however, they ended up selecting SVM as their top-performing classifier in a separate study [51]. Bian et al [54,55] and Afzal et al [36] compared the performance of multiple classifiers (SVM, NB, DT, k-nearest neighbors, random forest, multilayer perceptron) and selected the best performing for their experiment in the context of the same study (NB, DT, and SVM, respectively). We refer to the classifier that was selected for the classification task as the main classifier.

From the included articles, SVM was the most used classifier. Five studies used an SVM algorithm as one of their main experiment classifiers (Table 3), 2 studies used a neural network as their main classifier; Del Fiol et al [35] used a convolutional neural network (CNN), while Afzal et al [56] used a multilayer feed-forward artificial neural network (ANN). DT algorithms were used in 2 studies for their main text classification function [52,55]. Four of the included studies applied multiple classifying approaches [36,49,50,53].

Table 3. Types of machine learning classifiers used in the main experiment to assess performance in each of the included studies.

Author	Naïve Bayes	SVM[a]	Decision tree	Ensemble		Neural network
				Boosting	Stacking	
Aphinyanaphongs et al [49]	✓[b]	✓	N/A[c]	✓	N/A	N/A
Aphinyanaphongs et al [50]	✓	✓	N/A	✓	N/A	N/A
Aphinyanaphongs et al [51]	N/A	✓	N/A	N/A	N/A	N/A
Kilicoglu et al [53]	✓	✓	N/A	✓	✓	N/A
Lin et al [52]	N/A	N/A	✓	N/A	N/A	N/A
Afzal et al [36]	N/A	✓	N/A	N/A	N/A	N/A
Bian et al [54]	✓	N/A	N/A	N/A	N/A	N/A
Del Fiol et al [35]	N/A	N/A	N/A	N/A	N/A	✓
Bian et al [55]	N/A	N/A	✓	N/A	N/A	N/A
Afzal et al [56]	N/A	N/A	N/A	N/A	N/A	✓

[a] SVM: support vector machine.

[b] Applied.

[c] Not applied.

Comparator for Evaluating the Performance of the Classifiers

As per our inclusion criteria, to evaluate the performance of the machine learning method to classify articles appropriately, articles had to report a comparison of their applied machine learning model to a gold standard method such as gold standard high-quality articles retrieval method, for example, search filters, a manually annotated high-quality articles' dataset, or a baseline machine learning model for high-quality articles retrieval (Table 4). Aphinyanaphongs et al [49-51] used Clinical Query filters with sensitivity and specificity optimization [57]. The experiment conducted by Kilicoglu et al [53] evaluated their machine learning approach by applying it in a new dataset annotated by experts. The NB high-quality algorithm by Kilicoglu et al [53] was considered a comparator on its own for its high recall and was used as such by Bian and colleagues [54,55], who also used PubMed's best match as a comparator. Lin et al [52] used accuracy and k-value performance metrics in comparison to the results of the critical appraisal process by experts in the field. Also, Lin et al [52] has applied a comparison between their classifier, which was a DT, to other known text classifiers like SVM and ANN. Afzal et al [36] have used a SVM model for quality articles retrieval and compared its performance to the SVM model proposed by Sarker et al [58], reporting that their classifier achieved a higher performance with their reported features selected.

Del Fiol et al [35] was the first study to incorporate the use of deep learning in quality articles retrieval, relying on a CNN. Del Fiol and colleagues [35] compared their proposed classifier to the PubMed Clinical Queries broad filter since it achieves a nearly perfect recall. Also, they compared their proposed model to McMaster textword search and McMaster balanced search filter created by the Clinical Hedges group to evaluate the capabilities of their model of retrieving recently published evidence and achieving a balance between recall and precision [35]. Afzal et al [36], in their experiment using ANN, compared their model's results to the CNN results of Del Fiol et al [35], the DT results of Bian et al [55], and their prior experiment using an SVM for quality articles retrieval [35,56]. Also, Afzal et al [56] compared their proposed ANN to well-known algorithms used in the literature like NB, SVM, DT, and gradient boosted trees.

Table 4. The gold standard comparator used for evaluating machine learning models in the included studies.

Author	Comparator
Aphinyanaphongs et al [49-51]	• PubMed Clinical Query filter [57]
Kilicoglu et al [53]	• Testing dataset of 2000 articles annotated by experts (held-out testing dataset to test model's generalization)
Lin et al [52]	• Critical appraisal by domain expert • SVM[a] • Artificial neural network
Afzal et al [36]	• SVM proposed in Sarker et al [58]
Bian et al [54]	• Kilicoglu [53] high-quality classifier • PubMed's relevance sort
Del Fiol et al [35]	• PubMed Clinical Query filter • McMaster textword search • McMaster balanced filter
Bian et al [55]	• Kilicoglu et al [53] high-quality classifier • PubMed relevance sort • High-impact classifier with time-sensitive features included by Bian et al [54]
Afzal et al [56]	• Well-known algorithms used in the literature: NB[b], SVM, DT[c], GBT[d] • Models from past research by Del Fiol et al [35], Afzal et al [36], and Bian et al [55]

[a] SVM: support vector machine.

[b] NB: naïve Bayes.

[c] DT: decision tree.

[d] GBT: gradient boosted trees.

Performance Metrics

All included articles applied a supervised machine learning model. Validation by applying a resampling k-fold approach was used in 7 studies. Five used 10-fold cross-validation [35,36,49,52,53], and 2 studies relied on 5-fold cross-validation [50,51]. The most common performance metrics used in the included studies were sensitivity (recall), specificity, accuracy, area under the curve (AUC), F-measure, and precision (Table 5). The recall was generally high, above 85%, across all experiment classifiers except the SVM by Kilicoglu et al [53], and the NB and DT reported by Bian et al [54] and Bian et al [55], respectively, as both had a recall below 30%. Precision ranged from 9% to 86%, with the neural network of Afzal et al [56] and the SVM by Kilicoglu et al [53] the highest. AUC was measured in all studies and ranged from 0.73 to 0.99. Lin et al [52] and Bian et al [54,55] used novel performance metrics in their approaches. In the 2 studies by Bian and colleagues [54,55], performance was primarily determined by calculating the top 20 precision which is the measure of the percentage of true positive citations among the first 20 retrieved citations. Lin et al [52] used Cohen kappa (k-value) as their performance metric, which is the agreement between machine performance (observed value) and gold standard (expected value) [59,60]. Bian et al [54,55] reported a top 20 precision of 34% with their 2017 NB classifier and 24% in their 2019 experiment using a DT classifier. Lin et al [52] reported a k-value of 0.78 in their experiment.

Table 5. Highest reported performance characteristics of the main classifier algorithms reported in the included studies.

Classifier and author	Recall[a]	Specificity[b]	Precision[c]	F-score[d]	AUC[e]	Accuracy[f]
Support vector machine						
Aphinyanaphongs et al [49]	0.967	0.87	0.169	0.29[g]	0.98	0.893
Aphinyanaphongs et al [50]	0.96	0.86	0.18	0.30[g]	0.97	NR[h]
Aphinyanaphongs et al [51]	0.98	0.88	0.305	0.47[g]	0.95	NR
Kilicoglu et al [53]	0.229	NR	0.865	0.36	0.96	NR
Afzal et al [36]	NR	NR	NR	0.87	0.73	0.785
Naïve Bayes						
Aphinyanaphongs et al [49]	0.967	0.76	0.091	0.17[g]	0.95	0.787
Aphinyanaphongs et al [50]	NR	NR	NR	NR	0.95	NR
Kilicoglu et al [53]	0.975	NR	0.138	0.24	0.82	NR
Bian et al [54]	0.23	NR	0.33	0.21	NR	NR
Boosting						
Aphinyanaphongs et al [49]	0.967	0.786	0.099	0.18[g]	0.96	0.804
Aphinyanaphongs et al [50]	NR	NR	NR	NR	0.94	NR
Kilicoglu et al [53]	0.729	NR	0.823	0.77	0.97	NR
Neural network						
Del Fiol et al [35]	0.969	NR	0.346	0.51	NR	NR
Afzal et al [56]	0.951	NR	0.863	0.9	0.99	0.973
Decision tree						
Lin et al [52]	NR	NR	NR	NR	NR	0.854
Bian et al [55]	0.09	NR	0.39	0.14	NR	NR
Stacking						
Kilicoglu et al [53]	0.864	NR	0.747	0.801	0.98	NR

[a] Recall: proportion of correctly identified positives among the real positive.

[b] Specificity: the proportion of actual negatives, which got predicted as the negative (or true negative).

[c] Precision: proportion of correctly identified positives among all classified positives.

[d] F-score: harmonic mean of the precision and recall. F-score is equivalent to F1-score and used interchangeably.

[e] AUC: area under the curve traced out by graphing the true positive rate against the false positive rate. The higher the AUC, the better the classifier prediction.

[f] Accuracy: number of correctly predicted documents out of all classified documents.

[g] Calculated as F-measure=(2*precision*recall)/(precision+recall) using recall and precision when available from the articles.

[h] NR: not reported.

Discussion

Summary

To our knowledge, this is the first systematic review of machine learning approaches used to classify scientifically sound and clinically relevant studies from the biomedical literature. All included studies followed a supervised machine learning technique in which the learning algorithm depends on prelabeled data provided for training [41]. Despite the technological advancements from 2003 to 2020 when the studies were published, none reported applying unsupervised or active learning approaches for the classification of articles based on quality. Active learning is a subtype of machine learning in which the learning algorithm is allowed to select the data from which it learns by querying a human operator and can achieve a performance comparable to the standard supervised learning algorithms with fewer labeled data [21]. For example, active learning was used in the recent work by Gates et al [61] and Tsou et al [62], who used Abstrackr, a freely available active machine learning tool that automates the screening of titles and abstracts [63]. Abstrackr achieved 100% sensitivity after screening only 31.8% of the citations in the dataset [63].

There is a limited range of quality standards comprising the prelabeled training datasets across the included articles. ACP Journal Club and the Clinical Hedges follow the same inclusion and exclusion criteria for high-quality evidence [4].

Aphinyanaphongs et al [49] considered an article as high-quality if it were included in ACP Journal Club but considered only those classified as treatment, which limits their results to RCTs. The authors expanded their inclusion to articles tagged as treatment, diagnosis, prognosis, and etiology in their subsequent studies [50,51]. Having consistency across gold standard databases in classifier development strengthens our ability to compare performance. There are, however, limited manually annotated datasets available as these are time consuming and expensive to develop and require consistency and highly skilled people. Using studies that are included in guidelines and systematic reviews, as done by Bian et al [54,55] and Afzal et al [56], leverages screening work that has already been completed to a high standard; however, citations in guidelines may include lower quality evidence in the training process [64].

The limited availability of high-quality dataset options was highlighted by Afzal et al [56], and finding the ideal gold standard training dataset was the most reported limitation in the included studies. In our opinion, the ideal gold standard training dataset should cover some criteria to overcome the limitations reported in the articles. First, the gold standard should be defined by precise criteria for methodological rigor that is created and recognized by experts in the field [50]. Selection criteria for the gold standard should be unbiased. Aphinyanaphongs et al [50] described their concern toward the possibility of a selection bias by the ACP Journal Club editors in a particular year toward a certain topic. Second, the gold standard training dataset should cover a large enough sample of the high-quality class to properly train the model and overcome the class imbalance bias toward the majority class of studies that are not of high quality [63,65]. Third, the gold standard training set should cover multiple health care domains, as Lin et al [52] reported their high-quality dataset was limited only to cardiovascular diseases and would not perform as well if applied to another medical domain. Fourth, the gold standard training dataset should be up to date as much as possible, which was a limitation reported in both studies by Bian et al [54] and Afzal et al [56].

Another possible constraint affecting accurate prediction is the feature selection process. Del Fiol et al [35] stated that using MeSH-based features instead of the sole reliance on text features in their experiment could have improved the precision of their neural network. In consensus with the recommendation of Del Fiol et al [35], some of the included studies provided evidence that the use of a combination of features improves the overall performance of the classifiers. For example, in the experiment by Afzal et al [36], the combination of publication type and MeSH term features in addition to title and abstract features produced the best and the most stable results. Also, Kilicoglu et al [53] proved that the incorporation of MEDLINE citation metadata and Unified Medical Language System features in addition to words of titles and abstracts yielded the best performance. Such important features may not be immediately available at the time of indexing in MEDLINE [19], which poses a challenge in identifying recently published evidence [52,54].

There was a higher rate of incorporating SVM algorithms in the experiments by the study authors. SVMs are known for their high accuracy [66] and their low classification error [41], making them ideal for linear classification. Afzal et al [56] developed an ANN algorithm that had higher accuracy when compared with their previous SVM classifier [36]. Further applications of newer machine learning approaches will advance the knowledge base on these quickly evolving methods. While SVMs currently have good accuracy and low error rates, emerging approaches may well outperform them.

The main purpose of using machine learning in the classification of high-quality articles is to decrease the workload on those performing manual classification without losing relevant articles in the process. Recall, the proportion of correctly identified high-quality articles from the high-quality pool, is the most important metric to be used, followed by precision, the proportion of correctly identified positive articles among all those classified as positive. The included studies reported a range of recall and precision some of which would not meet the objective of identifying the high-quality articles correctly. For example, the NB classifier developed by Bian et al [54] performed significantly less than the NB by Kilicoglu et al [53] and PubMed Best Match in terms of recall (23% vs 55% and 65%, respectively). Despite performing worse in recall, their classifier achieved a higher precision (33% vs 5% and 4%) [54].

Additionally, accuracy, the number of correctly predicted documents out of all classified documents, is considered a common metric for evaluating classifiers; however, its use is considered inappropriate to evaluate imbalanced dataset classification [67]. For example, a classifier labeling all entries as false (given that false is the majority class) would have high accuracy but would fail to perform the needed task of accurately classifying the passing articles (rare class), making it useless [68]. The harmonic mean of the recall and precision measurements is the F-score, and it is used to evaluate the machine learning algorithms implemented on unbalanced datasets [67]. F-score was first used in the study by Kilicoglu et al [53] where the performance of the classifiers was reported using recall, precision, F-score, and AUC, without including accuracy. Additionally, Afzal et al [36] did not rely on recall to compare between multiple classifiers; instead, they used the F-score, precision, and accuracy. Also, they have applied a novel approach to compare between the classifiers, in which they summed the metrics for a classifier with a higher sum reflecting better performance [36].

The highest reported recall in our review was 98% with the SVM developed by Aphinyanaphongs and Aliferis [51], however, the algorithm had low precision of 30.5%. The best balance between recall and precision was achieved by the ANN approach used by Afzal et al [56], which reported a high recall of 95.1% and a high precision of 86.3%, thereby achieving the target of not losing quality literature while decreasing the manual classification workload.

The experiment by Kilicoglu et al [53] assesses the effect of applying 3 different machine learning classifiers (SVM, NB, boosting, and ensemble) trained using the same Clinical Hedges dataset on the overall performance of the resulting models. Using multiple feature set combinations, the highest recall was achieved by the NB classifier, and the highest F-scores were achieved by ensemble (0.80) and text-boosting (0.77) based models [53]. Only the studies by Aphinyanaphongs and colleagues [49,50] and Kilicoglu et al [53] incorporated ensemble techniques in the development of their main classifiers, and their results suggest that using multiple classifiers in combination can improve the balance between recall and precision (the F-score).

Strengths and Limitations

This is the first systematic review to characterize the machine learning approaches in high-quality article retrieval. When narrowing our research question, we excluded other text summarization and text categorization approaches being used in the biomedical literature. These include but are not limited to studies concerned with the automation of the systematic review process [69,70], biomedical literature summarization [71], and semantic models’ applications in the biomedical literature [72]. Given the technical nature of the application of machine learning approaches for text classification, we expanded our search beyond clinical bibliographic databases to include those which index technical articles.

Across the included studies, some steps were not fully reported in the methods, including preprocessing steps, cross-validation folds, and features selected. To our knowledge, there are no reporting guidelines for machine learning approaches being applied for literature retrieval. The Equator Network includes 6 reporting guidelines for machine learning approaches; however, all 6 are focused on articles applying machine learning in clinical settings [73]. For example, the most recently published guideline focuses on the reporting of interventions involving artificial intelligence in clinical trial protocols [74]. The lack of reporting guidance for the NLP component of machine learning being applied in the biomedical literature creates a noticeable gap in reporting the steps of the applied approach, features used and justification for their use, and inconsistency in the reported performance achieved by the machine. As a result, there was a lack of consistency in the reporting of results and methods provided by the authors, which also limits our ability to compare the performance of the classifiers. Also, one of the limitations developing the review was the inability to directly compare the performance of the models across the included studies because of the different training datasets and the applied settings. Finally, a challenge with machine learning is that the algorithms are considered as being derived in a black box; an enigmatic interpretation that the machines provide findings and predictions without any accompanying explanation [75].

Conclusion

Despite the longevity of research for the identification of high-quality literature using machine learning, evidence is still scarce and slowly progressing over time, and determining the most reliable approach is difficult as the field is quickly evolving. This slow progression in the field may have been caused by the lack of publicly available standard benchmarks for the identification of high-quality articles biomedical literature to compare the performance of the proposed methods. A similar problem was addressed in the molecular machine learning domain by creating MolecularNet, a large-scale, open-source, and high-quality benchmark for molecular learning algorithms [76]. Our review provides a summary of current approaches and performance of machine learning models applied to retrieve high-quality evidence for clinical consideration from the biomedical literature and highlights the importance of selecting optimal quality gold standard data for training. The findings include that the use of different feature sets in combination with text features is likely to improve the performance of machine learning models. There is a lack of reporting consistency in the literature which makes replication of the experiments difficult. Supervised machine learning has been the focus to date. The rapid development in the field of NLP and the availability of new state of the art techniques such as Bidirectional Encoder Representations from Transformers (BERT) for language understanding [77] and bio-BERT for biomedical text mining [78] hold promise for future advances in the field of information extraction from the biomedical literature. Considering the increasingly available data to apply these approaches to, we anticipate that the performance of classifiers to identify high-quality evidence will continue to grow.

Acknowledgments

We thank Rita Jezrawi and Hamza Issa for assistance screening articles for inclusion. We thank Dr Mehrdad Roham for proofreading the final version of the manuscript.

Authors' Contributions

All authors contributed to the design of the study. TN and WA developed the search strategies. WA ran the searches and led the screening and data abstraction. All authors contributed to the interpretation of the data. WA and CL drafted early versions of the manuscript. All authors supervised the study and reviewed and provided revisions to the manuscript. All authors approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Features selected by the included articles.
[PDF File (Adobe PDF File), 187 KB - medinform_v9i9e30401_app1.pdf]

References

1. Guyatt G, Jaeschke R, Wilson M, Montori V, Richardson W. What Is evidence-based medicine? In: Guyatt G, Rennie D, Meade MO, Cook DJ, editors. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3rd Edition. New York: McGraw-Hill Education; 2015:7-14.
2. Kamath S, Guyatt G. Importance of evidence-based medicine on research and practice. Indian J Anaesth 2016 Sep;60(9):622-625 [FREE Full text] [doi: 10.4103/0019-5049.190615] [Medline: 27729686]
3. MEDLINE PubMed Production Statistics. URL: https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html [accessed 2021-08-23]
4. Wilczynski NL, Morgan D, Haynes RB, Hedges Team. An overview of the design and methods for retrieving high-quality studies for clinical care. BMC Med Inform Decis Mak 2005 Jun 21;5:20 [FREE Full text] [doi: 10.1186/1472-6947-5-20] [Medline: 15969765]
5. Beale S, Duffy S, Glanville J, Lefebvre C, Wright D, McCool R, et al. Choosing and using methodological search filters: searchers' views. Health Info Libr J 2014 Apr 23;31(2):133-147. [doi: 10.1111/hir.12062]
6. Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically sound causation studies in MEDLINE. AMIA Annu Symp Proc 2003:719-723 [FREE Full text] [Medline: 14728267]
7. Leeflang M, Scholten R, Rutjes A, Reitsma J, Bossuyt P. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006 Mar;59(3):234-240. [doi: 10.1016/j.jclinepi.2005.07.014] [Medline: 16488353]
8. Damarell RA, May N, Hammond S, Sladek RM, Tieman JJ. Topic search filters: a systematic scoping review. Health Info Libr J 2019 Mar 21;36(1):4-40. [doi: 10.1111/hir.12244] [Medline: 30578606]
9. McKibbon K, Wilczynski NL, Haynes R, Hedges Team. Retrieving randomized controlled trials from medline: a comparison of 38 published search filters. Health Info Libr J 2009 Sep;26(3):187-202 [FREE Full text] [doi: 10.1111/j.1471-1842.2008.00827.x] [Medline: 19712211]
10. Miller PA, McKibbon KA, Haynes RB. A quantitative analysis of research publications in physical therapy journals. Phys Ther 2003 Feb;83(2):123-131. [Medline: 12564948]
11. Search filters for MEDLINE in Ovid Syntax and the PubMed translation. Health Information Research Unit, McMaster University. URL: https://hiru.mcmaster.ca/hiru/HIRU_Hedges_MEDLINE_Strategies.aspx [accessed 2021-01-31]
12. Hedges. Health Information Research Unit, McMaster University. URL: https://hiru.mcmaster.ca/hiru/HIRU_Hedges_home.aspx [accessed 2021-05-31]
13. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996 Feb;17(1):1-12. [doi: 10.1016/0197-2456(95)00134-4] [Medline: 8721797]
14. Higgins JPT, Altman DG, Gøtzsche PC, Jüni P, Moher D, Oxman AD, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. BMJ 2011 Oct 18;343(oct18 2):d5928-d5928 [FREE Full text] [doi: 10.1136/bmj.d5928] [Medline: 22008217]
15. Inclusion Criteria. Health Information Research Unit, McMaster University. 2019. URL: https://hiru.mcmaster.ca/hiru/InclusionCriteria.html [accessed 2021-05-31]
16. Wong SS, Wilczynski NL, Haynes RB, Hedges Team. Developing optimal search strategies for detecting clinically relevant qualitative studies in MEDLINE. Stud Health Technol Inform 2004;107(Pt 1):311-316. [Medline: 15360825]
17. Gill PJ, Roberts NW, Wang KY, Heneghan C. Development of a search filter for identifying studies completed in primary care. Fam Pract 2014 Dec 18;31(6):739-745. [doi: 10.1093/fampra/cmu066] [Medline: 25326923]
18. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc 2013;20(2):363-368 [FREE Full text] [doi: 10.1136/amiajnl-2012-001075] [Medline: 23019242]
19. Irwin AN, Rackham D. Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. Res Social Adm Pharm 2017;13(2):389-393. [doi: 10.1016/j.sapharm.2016.04.006] [Medline: 27215603]
20. Tsafnat G, Glasziou P, Karystianis G, Coiera E. Automated screening of research studies for systematic reviews using study characteristics. Syst Rev 2018 Apr 25;7(1):64 [FREE Full text] [doi: 10.1186/s13643-018-0724-7] [Medline: 29695296]
21. Mohri M, Rostamizadeh A, Talwalkar A. In: Bach F, editor. Foundations of Machine Learning, Second edition. Cambridge: MIT Press; 2018.
22. Mitchell T. Machine Learning, 1st ed. New York: McGraw-Hill Science; 1997:432.
23. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag 2005;19(2):64-72. [Medline: 15869215]
24. Bahri S, Zoghlami N, Abed M, Tavares JMRS. Big data for healthcare: a survey. IEEE Access 2019;7:7397-7408. [doi: 10.1109/access.2018.2889180]
25. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018 Apr 03;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]
26. Wiens J, Shenoy ES. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clin Infect Dis 2018 Jan 06;66(1):149-153 [FREE Full text] [doi: 10.1093/cid/cix731] [Medline: 29020316]

27. Wiens J, Guttag J, Horvitz E. Patient risk stratification with time-varying parameters: a multitask learning approach. J Mach Learning Res 2016;17:1-23 [FREE Full text] [doi: 10.1016/b978-0-12-802121-7.00045-5]
28. Wiens J, Guttag J, Horvitz E. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. J Am Med Inform Assoc 2014;21(4):699-706 [FREE Full text] [doi: 10.1136/amiajnl-2013-002162] [Medline: 24481703]
29. Wiens J, Horvitz E, Guttag J. Patient risk stratification for hospital-associated C. diff as a time-series classification task. In: Pereira F, Burges C, Bottou L, Weinberger K, editors. Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc; 2012:467-475.
30. Burkov A. The Hundred-Page Machine Learning Book, 1st ed. Quebec City: Andriy Burkov; 2019.
31. Hearst M. Text Data Mining. In: Mitkov R, editor. The Oxford Handbook of Computational Linguistics (1st ed). Oxford: Oxford University Press; 2012.
32. Gong L. Application of biomedical text mining. In: Artificial Intelligence—Emerging Trends and Applications InTech. London: IntechOpen; 2018.
33. Davidoff F, Miglus J. Delivering clinical evidence where it's needed: building an information system worthy of the profession. JAMA 2011 May 11;305(18):1906-1907. [doi: 10.1001/jama.2011.619] [Medline: 21558524]
34. Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, et al. Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 2014 Dec;52:457-467 [FREE Full text] [doi: 10.1016/j.jbi.2014.06.009] [Medline: 25016293]
35. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: comparative analytic study. J Med Internet Res 2018 Jun 25;20(6):e10281 [FREE Full text] [doi: 10.2196/10281] [Medline: 29941415]
36. Afzal M, Hussain M, Haynes RB, Lee S. Context-aware grading of quality evidences for evidence-based decision-making. Health Informatics J 2019 Jun;25(2):429-445 [FREE Full text] [doi: 10.1177/1460458217719560] [Medline: 28766402]
37. Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, et al. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. Genet Med 2012 Jul;14(7):663-669 [FREE Full text] [doi: 10.1038/gim.2012.7] [Medline: 22481134]
38. Wallace B, Small K, Brodley C, Lau J, Trikalinos T. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. Proc 2nd ACM SIGHIT Int Health Informat Symp 2012:819-824 [FREE Full text] [doi: 10.1145/2110363.2110464]
39. Rindflesch T, Fiszman M, Libbus B. Semantic interpretation for the biomedical research literature. Med Informatics 2006:399-422. [doi: 10.1007/0-387-25739-x_14]
40. Holzinger A. Machine learning for health informatics. LNCS 2016;9605:1-24. [doi: 10.1007/978-3-319-50478-0_1]
41. Dey A. Machine learning algorithms: a review. Int J Comput Sci Inf Technol 2016;7(3):1174-1179 [FREE Full text]
42. Cohen S. The basics of machine learning: strategies and techniques. In: Artificial Intelligence and Deep Learning in Pathology. Philadelphia: Elsevier; 2021:13-40.
43. Welling M. A First Encounter with Machine Learning. University of California Irvine. 2011. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.441.6238&rep=rep1&type=pdf [accessed 2021-08-23]
44. Wang S. Artificial neural network. In: Interdisciplinary Computing in Java Programming. Boston: Springer US; 2003:81-100.
45. Hiregoudar SB, Manjunath K, Patil KS. A survey: research summary on neural networks. Int J Res Engineer Technol 2014 May 25;03(15):385-389. [doi: 10.15623/ijret.2014.0315076]
46. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Res Synth Methods 2018 Dec;9(4):602-614 [FREE Full text] [doi: 10.1002/jrsm.1287] [Medline: 29314757]
47. Agarwal B, Mittal N. Text classification using machine learning methods: a survey. 2014 Presented at: Proceedings of the Second International Conference on Soft Computing for Problem Solving; 2014; New Delhi p. 701-709. [doi: 10.1007/978-81-322-1602-5_75]
48. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021 Mar 29;372:n71 [FREE Full text] [doi: 10.1136/bmj.n71] [Medline: 33782057]
49. Aphinyanaphongs Y, Aliferis CF. Text categorization models for retrieval of high quality articles in internal medicine. AMIA Annu Symp Proc 2003:31-35 [FREE Full text] [Medline: 14728128]
50. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc 2005;12(2):207-216 [FREE Full text] [doi: 10.1197/jamia.M1641] [Medline: 15561789]
51. Aphinyanaphongs Y, Aliferis C. Prospective validation of text categorization filters for identifying high-quality, content-specific articles in MEDLINE. AMIA Annu Symp Proc 2006:6-10 [FREE Full text] [Medline: 17238292]
52. Lin J, Chang C, Lin M, Ebell MH, Chiang J. Automating the process of critical appraisal and assessing the strength of evidence with information extraction technology. J Eval Clin Pract 2011 Aug;17(4):832-838. [doi: 10.1111/j.1365-2753.2011.01712.x] [Medline: 21707873]

53. Kilicoglu H, Demner-Fushman D, Rindflesch TC, Wilczynski NL, Haynes RB. Towards automatic recognition of scientifically rigorous clinical research evidence. J Am Med Inform Assoc 2009;16(1):25-31 [FREE Full text] [doi: 10.1197/jamia.M2996] [Medline: 18952929]
54. Bian J, Morid MA, Jonnalagadda S, Luo G, Del Fiol G. Automatic identification of high impact articles in PubMed to support clinical decision making. J Biomed Inform 2017 Sep;73:95-103 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.015] [Medline: 28756159]
55. Bian J, Abdelrahman S, Shi J, Del Fiol G. Automatic identification of recent high impact clinical articles in PubMed to support clinical decision making using time-agnostic features. J Biomed Inform 2019 Jan;89:1-10 [FREE Full text] [doi: 10.1016/j.jbi.2018.11.010] [Medline: 30468912]
56. Afzal M, Park BJ, Hussain M, Lee S. Deep learning based biomedical literature classification using criteria of scientific rigor. Electronics (Switzerland) 2020 Aug 05;9(8):1-12. [doi: 10.3390/electronics9081253]
57. PubMed Clinical Queries Search Filter. National Library of Medicine. URL: https://pubmed.ncbi.nlm.nih.gov/clinical/ [accessed 2021-05-01]
58. Sarker A, Mollá D, Paris C. Automatic evidence quality prediction to support evidence-based decision making. Artif Intell Med 2015 Jun;64(2):89-103. [doi: 10.1016/j.artmed.2015.04.001] [Medline: 25983133]
59. Chiang J, Lin J, Yang C. Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). J Am Med Inform Assoc 2010 May 01;17(3):245-252. [doi: 10.1136/jamia.2009.000182]
60. Rau G, Shih Y. Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data. J Engl Acad Purposes 2021 Sep;53:101026. [doi: 10.1016/j.jeap.2021.101026]
61. Gates A, Johnson C, Hartling L. Technology-assisted title and abstract screening for systematic reviews: a retrospective evaluation of the Abstrackr machine learning tool. Syst Rev 2018 Mar 12;7(1):45 [FREE Full text] [doi: 10.1186/s13643-018-0707-8] [Medline: 29530097]
62. Tsou AY, Treadwell JR, Erinoff E, Schoelles K. Machine learning for screening prioritization in systematic reviews: comparative performance of Abstrackr and EPPI-Reviewer. Syst Rev 2020 Apr 02;9(1):73 [FREE Full text] [doi: 10.1186/s13643-020-01324-7] [Medline: 32241297]
63. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinformatics 2010 Jan 26;11(1):55 [FREE Full text] [doi: 10.1186/1471-2105-11-55] [Medline: 20102628]
64. Venus C, Jamrozik E. Evidence-poor medicine: just how evidence-based are Australian clinical practice guidelines? Intern Med J 2020 Jan 14;50(1):30-37. [doi: 10.1111/imj.14466] [Medline: 31943616]
65. Lanera C, Berchialla P, Sharma A, Minto C, Gregori D, Baldi I. Screening PubMed abstracts: is class imbalance always a challenge to machine learning? Syst Rev 2019 Dec 06;8(1):317 [FREE Full text] [doi: 10.1186/s13643-019-1245-8] [Medline: 31810495]
66. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res 2014;15(1):3181.
67. Bekkar M, Djema H, Alitouche T. Evaluation measures for models assessment over imbalanced data sets. J Inf Engineer Applic 2013;3:27-38 [FREE Full text]
68. Tang Y, Zhang YQ, Chawla N, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Trans Syst Man Cybern B 2009 Feb;39(1):281-288. [doi: 10.1109/tsmcb.2008.2002909]
69. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev 2015 Jan 14;4:5 [FREE Full text] [doi: 10.1186/2046-4053-4-5] [Medline: 25588314]
70. Shakeel Y, Krüger J, Nostitz-Wallwitz IV, Saake G, Leich T. Automated selection and quality assessment of primary studies. J Data Inf Qual 2020 Jan 23;12(1):1-26. [doi: 10.1145/3356901]
71. Yoo I, Hu X, Song I. A coherent graph-based semantic clustering and summarization approach for biomedical literature and a new summarization evaluation method. BMC Bioinformatics 2007 Nov 27;8 Suppl 9:S4. [doi: 10.1186/1471-2105-8-S9-S4] [Medline: 18047705]
72. Arguello Casteleiro M, Maseda Fernandez D, Demetriou G, Read W, Fernandez Prieto MJ, Des Diz J, et al. A case study on sepsis using PubMed and deep learning for ontology learning. Stud Health Technol Inform 2017;235:516-520. [Medline: 28423846]
73. EQUATOR Network: Reporting Guidelines. URL: https://www.equator-network.org/reporting-guidelines/ [accessed 2021-08-24]
74. Rivera SC, Liu X, Chan A, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020 Sep 09;370:m3210 [FREE Full text] [doi: 10.1136/bmj.m3210] [Medline: 32907797]
75. Watson DS, Krutzinna J, Bruce IN, Griffiths CE, McInnes IB, Barnes MR, et al. Clinical applications of machine learning algorithms: beyond the black box. BMJ 2019 Mar 12;364:l886. [doi: 10.1136/bmj.l886] [Medline: 30862612]
76. Wu Z, Ramsundar B, Feinberg E, Gomes J, Geniesse C, Pappu AS, et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018 Jan 14;9(2):513-530 [FREE Full text] [doi: 10.1039/c7sc02664a] [Medline: 29629118]

77. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 10, 2018. [FREE Full text]

78. 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]

Abbreviations

AUC: area under the receiver operating characteristic curve
ACP: American College of Physicians
ANN: artificial neural network
BERT: Bidirectional Encoder Representations from Transformers
CNN: convolutional neural network
DT: decision tree
EBM: evidence-based medicine
MeSH: Medical Subject Heading
NB: naïve Bayes
NLP: natural language processing
RCT: randomized controlled trial
SVM: support vector machine

Edited by C Lovis; submitted 13.05.21; peer-reviewed by H Kalicoglu, M Afzal; comments to author 02.07.21; revised version received 15.07.21; accepted 25.07.21; published 09.09.21.

Please cite as:
Abdelkader W, Navarro T, Parrish R, Cotoi C, Germini F, Iorio A, Haynes RB, Lokker C
Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review
JMIR Med Inform 2021;9(9):e30401
URL: https://medinform.jmir.org/2021/9/e30401
doi:10.2196/30401
PMID:34499041

Corrigenda and Addenda

Correction: Predicting Health Material Accessibility: Development of Machine Learning Algorithms

Meng Ji[1], PhD; Yanmeng Liu[1], MA; Tianyong Hao[2], PhD

[1]School of Languages and Cultures, The University of Sydney, Sydney, Australia

[2]School of Computer Science, South China Normal University, Guangdong, China

Corresponding Author:
Tianyong Hao, PhD
School of Computer Science
South China Normal University
No.55 West Zhongshan Avenue, Shipai, Tianhe District
Guangdong, 510631
China
Phone: 86 15626239317
Email: haoty@m.scnu.edu.cn

Related Article:

Correction of: https://medinform.jmir.org/2021/9/e29175

(JMIR Med Inform 2021;9(9):e33385) doi:10.2196/33385

In “Predicting Health Material Accessibility: Development of Machine Learning Algorithms” (JMIR Med Inform 2021;9(9):e29175) the authors noted some errors. The following changes have been made to correct these errors:

Author Metadata

In the originally published paper, Affiliation 1 appeared as follows:

> *School of Languages and Culture, The University of Sydney, Sydney, Australia*

It is now corrected as follows:

> *School of Languages and Cultures, The University of Sydney, Sydney, Australia*

Abstract

- Under "Methods," the phrase "*We applied 10-fold cross-validation on the whole data set...*" has been replaced by "*We applied 5-fold cross-validation on the whole data set....*"
- Under "Results," the sentences "*The results showed that ensemble tree (LogitBoost) outperformed in terms of AUC (0.97), sensitivity (0.966), specificity (0.972), and accuracy (0.969). Decision tree (AUC 0.924, sensitivity 0.912, specificity 0.9358, and accuracy 0.924) and SVM (AUC 0.8946, sensitivity 0.8952, specificity 0.894, and accuracy 0.8946) followed closely. Decision tree, ensemble tree, and SVM achieved statistically significant improvement over logistic regression in AUC, specificity, and accuracy. As the best performing algorithm, ensemble tree reached statistically significant improvement over SVM in AUC, specificity, and accuracy, and statistically significant improvement over decision tree in sensitivity*" have been replaced by "*The results showed that ensemble classifier (LogitBoost) outperformed in terms of AUC (0.858), sensitivity (0.787), specificity (0.813), and accuracy (0.802). Support vector machine (AUC 0.848, sensitivity 0.783, specificity 0.791, and accuracy 0.786) and decision tree (AUC 0.754, sensitivity 0.7174, specificity 0.7424, and accuracy 0.732) followed. Ensemble classifier (LogitBoost), support vector machine, and decision tree achieved statistically significant improvement over logistic regression in AUC, sensitivity, specificity, and accuracy. Support vector machine reached statistically significant improvement over decision tree in AUC and accuracy. As the best performing algorithm, ensemble classifier (LogitBoost) reached statistically significant improvement over decision tree in AUC, sensitivity, specificity, and accuracy.*"

Introduction

- Under "Material Collection and Classification," the last sentence "*The final classification contained two sets of texts: easy (n=499) versus difficult (n=501;...*" has been replaced by "*The final classification contained two sets of texts: easy (n=495) versus difficult (n=505;....*"
- Under "Material Annotation and Semantic Feature Extraction," the sentence "*With USAS, we collected 108 semantic features*" has been replaced by "*With USAS, we collected 113 semantic features.*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*A total of 29 of the 113 semantic features were identified as statistically significant…*" has

been replaced by "*A total of 26 of the 113 semantic features were identified as statistically significant....*"

- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*The mean score of Z8 in health texts of higher understandability was 52.91, this dropped to 20.15 ...*" has been replaced by "*The mean score of Z8 in health texts of higher understandability was 52.84, this dropped to 20.48....*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*…was 0.929 (95% CI 0.905-0.953)...easy reading was 0.929...*" has been replaced by "*...was 0.928 (95% CI 0.905-0.951)… easy reading was 0.928....*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*…were identified as statistically significant (P=.005)*" has been replaced by "*...were identified as statistically significant (P=.01).*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*The odds ratio of Z7 was 0.845 (95% CI 0.751-0.951),… a difficult text was 84.5%...*" has been replaced by "*The odds ratio of Z7 was 0.86 (95% CI 0.767-0.964), …a difficult text was 86%....*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the first paragraph, the sentence "*The large semantic category X2 (mental actions and process) was detected as a large contributor to the cognitive accessibility of health texts (odds ratio Exp(B) 0.92, 95% CI 0.852-0.995; P=.04). Typical expressions included in the X2 class were English expressions related to reasoning and thinking and levels of belief or skepticism. Terms of knowledge acquisition, perception, and retrospection were included in this broad category, such as familiarize, forget, reflect, or become aware*" has been deleted.
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the first sentence "*The logistic regression result (Multimedia Appendix 1) also identified 13 semantic features...*" has been replaced by "*The logistic regression result (Multimedia Appendix 1) also identified 12 semantic features....*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the sentence "*Typical examples were B3 (medicines and medical treatment; odds ratio Exp(B) 1.042, 95% CI 1.012-1.073; P=.005), Z99 (out-of-dictionary words; odds ratio Exp(B) 1.01, 95% CI 1.004-1.017; P=.003), L2 (living creatures: animals, microorganism, virus, bacteria, etc; odds ratio Exp(B) 1.082, 95% CI 1.003-1.167; P=.04), and W5 (environmental terms: pollutants, carcinogens, inhalable particles, etc; odds ratio Exp(B) 2.244, 95% CI 1.11-4.538; P=.02)*" has been replaced by "*Typical examples were B3 (medicines and medical treatment; odds ratio Exp(B) 1.041, 95% CI 1.012-1.071; P=.005), Z99 (out-of-dictionary words; odds ratio Exp(B) 1.011, 95% CI 1.004-1.018; P=.001), L2 (living creatures: animals, microorganism, virus, bacteria, etc; odds ratio Exp(B) 1.080, 95% CI 1.005-1.162; P=.036), and W5 (environmental terms: pollutants, carcinogens, inhalable particles, etc.; odds ratio Exp(B) 2.441, 95% CI 1.173-5.077; P=.017).*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the sentence "*For example, the relatively large odds ratios (mean 2.244, 95% CI 1.11-4.538) of W5 encompassing terms related to environmental exposure and health risks indicates that, with the increase of one word in this particular category, the odds of a health text being a difficult text over the odds of the text being an easy text for the target readers was 2.244, or in terms of percentage change, this represents an increase of 124.4% of the text from an easy text to a very difficult health reading*" has been replaced by "*For example, the relatively large odds ratios (2.441, 95% CI 1.173-5.077) of W5 encompassing terms related to environmental exposure and health risks indicates that, with the increase of one word in this particular category, the odds of a health text being a difficult text over the odds of the text being an easy text for the target readers was 2.441, or in terms of percentage change, this represents an increase of 144.1% of the text from an easy text to a very difficult health reading.*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the sentence "*To a lesser extent, the odds ratio of 1.082 of L2 (living creatures including microorganisms) indicates that with the increase of one word in this class, the perceived difficulty level (hard-to-understand class) of the health text increased by a mean 8.2% (95% CI 0.3%-16.7%) depending on the vocabulary range of English health terms of the readers*" has been replaced by "*To a lesser extent, the odds ratio of 1.080 of L2 (living creatures including microorganisms) indicates that with the increase of one word in this class, the perceived difficulty level (hard-to-understand class) of the health text increased by a mean 8.0% (95% CI 0.5%-16.2%) depending on the vocabulary range of English health terms of the readers.*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the sentence "*These include A2 (general or abstract terms denoting the propensity for changes, such as adapt, adjust for, conversion, and alter; odds ratio 1.057, 95% CI 1.005-1.111; P=.03), A7 (abstract terms of modality, such as possibility, necessity, and certainty; odds ratio 1.099, 95% CI 1.006-1.2; P=.04), A11 (abstract terms denoting importance, significance, noticeability, or markedness; odds ratio 1.164, 95% CI 1.003-1.351; P=.045)*" has been replaced by "*These include A11 (abstract terms denoting importance, significance, noticeability, or markedness; odds ratio 1.219, 95% CI 1.070-1.388; P=.003).*"
- Under "Statistical Analysis of Multidimensional Semantic Features in English Educational Health Texts," in the second paragraph, the sentence "*This means that with the increase of one word in the A11 class, the odds of the health text*

being seen as a hard-to-understand text over the text being seen as an easy text was 1.164, or an increase of 16.4%" has been replaced by "*This means that with the increase of one unit in the A11 class, the odds of the health text being seen as a hard-to-understand text over the text being seen as an easy text was 1.219, or an increase of 21.9%*."

Methods

- Under "Methods," in the second paragraph, the sentences " *For a decision tree classifier, the best-point hyperparameters (Figure 1) were the maximum number of tree splits (n=22) based on Gini diversity index (minimum parent node size n=10). The observed minimal classification error of the optimized decision tree model was 0.203. For an ensemble classifier, the best-point hyperparameters (Figure 2) reached an observed minimum classification error of 0.14091. The optimized hyperparameters were the ensemble method (LogitBoost), number of learners (n=302), learning rate (0.15456), and maximum number of splits (n=9). For SVM, the best-point hyperparameters (Figure 3) were box constraint level (0.014832; kernel function: linear). The observed minimum classification error was 0.18722, lower than the optimized decision tree model (0.203) but higher than the optimized ensemble classifier (0.14091)*" have been replaced by " *For a decision tree classifier, the best-point hyperparameters (Figure 1) were the maximum number of tree splits (n=22) based on maximum deviance reduction. The observed minimal classification error of the optimized decision tree model was 0.215. For an ensemble classifier, the best-point hyperparameters (Figure 2) reached an observed minimum classification error of 0.168. The optimized hyperparameters were the ensemble method (LogitBoost), number of learners (n=210), learning rate (0.1), and maximum number of splits (n=22). For SVM, the best-point hyperparameters (Figure 3) were box constraint level (0.1), kernel function (cubic). The observed minimum classification error was 0.1944, lower than the optimized decision tree model (with a difference of 0.0206) but higher than the optimized ensemble classifier (with a difference of 0.0264).*"

Results

- Under "Results," in the first paragraph, the sentences "*The mean scores and SDs of the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy were obtained through 10-fold cross-validation. The cross-validation divided the entire data set into 10 folds of equal size. In each iteration, 9 folds were used for the training data, and the remaining fold was used as the testing data. As a result, on completion of the 10-fold cross-validation, each fold was used as the testing data exactly once. We used pairwise corrected resampled t test to counteract the issue of multiple comparisons. As the result, the significance level was adjusted to .008 (n=6; α=.05) using Bonferroni correction*" have been replaced by "*The mean scores and standard deviations of the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy were obtained through 5-fold cross-validation. The cross-validation divided the entire data set into 5 folds of equal size. In each iteration, 4 folds were used for the training data, and the remaining fold was used as the testing data. As a result, on completion of the 5-fold cross-validation, each fold was used as the testing data exactly once. We used paired-sample comparisons to investigate the area under the operating characteristic curve (AUC), sensitivity, specificity, and accuracy differences of four machine learning algorithms (n=6; α=.05).*"
- Under "Results," the second paragraph " *Table 2 shows that, in terms of AUC, ensemble classifier (LogitBoost), decision tree, and SVM reached statistically improved AUC over logistic regression (0.802): LogitBoost (0.97; P<.001), decision tree (0.924; P<.001), and SVM (0.8946, P=.002). In terms of sensitivity, only LogitBoost (0.966; P<.001) reached statistically significant improvement over logistic regression (0.8364), whereas decision tree (0.9122) and SVM (0.8952) had similar sensitivity as logistic regression. In terms of model specificity, LogitBoost, decision tree, and SVM all reached statistically improved specificity over logistic regression (0.7694): LogitBoost (0.972; P=.002), decision tree (0.9358; P=.003), and SVM (0.894; P=.004). Lastly, with regard to model overall accuracy, again, LogitBoost, decision tree, and SVM outperformed logistic regression (0.8029): LogitBoost (0.969; P<.001), decision tree (0.924; P<.001), and SVM (0.8946; P=.002). Comparing LogitBoost, decision tree, and SVM, the former two algorithms outperformed SVM consistently in AUC (P=.001), sensitivity (P=.007), and accuracy (P=.001), and LogitBoost and SVM outperformed decision tree in terms of model specificity (P=.003), using the adjusted .008 as the significance level of paired-sample comparisons (Bonferroni correction: n=6; α=.05). These results suggest that, when using semantic features as predictor variables, the most stable and highest-performing algorithm is ensemble classifier (LogitBoost), followed by optimized decision tree. LogitBoost, decision tree, and SVM all achieved statistically significant improvement over logistic regression in AUC, specificity, and accuracy. Decision tree and SVM did not improve over logistic regression in terms of sensitivity, but LogitBoost did. Overall, the best AUC, sensitivity, specificity, and accuracy were achieved by LogitBoost as an ensemble classifier (Figure 4)*" has been replaced by " *Table 2 shows that, in terms of AUC, ensemble classifier (LogitBoost), decision tree, and SVM reached statistically improved AUC over logistic regression (0.614): ensemble classifier (0.858; P=.001), decision tree (0.754; P=.004), and SVM (0.848, P=.001). In terms of sensitivity (Table 3), ensemble classifier (0.787, P=.020), decision tree (0.7174, P=.036), and SVM (0.783; P<.001) reached statistically significant improvement over logistic regression (0.6282). In terms of model specificity (Table 4), ensemble classifier, decision tree, and SVM all reached statistically improved specificity over logistic regression (0.5724): ensemble classifier (0.813; P=.001), decision tree (0.7424; P=.009), and SVM (0.791; P=.007). Lastly, with regard to model overall accuracy (Table 5), again, LogitBoost, decision tree, and SVM outperformed logistic regression*

(0.601): ensemble classifier (0.802; P=.001), decision tree (0.732; P=.003), and SVM (0.786; P=.001). Comparing SVM, ensemble classifier and decision tree, the former two algorithms outperformed decision tree consistently in AUC (P=.001 and P<.001, respectively), and accuracy (P=.022 and P=.001, respectively). Only ensemble classifier outperformed decision tree significantly in terms of model sensitivity (P=.024), and specificity (P=.010), using the paired-sample comparisons (n=6; α=.05). These results suggest that, when using semantic features as predictor variables, the most stable and highest-performing algorithm is ensemble classifier (LogitBoost), followed by SVM. Ensemble classifier, decision tree, and SVM all achieved statistically significant improvement over logistic regression in AUC, specificity, sensitivity, and accuracy. SVM did not improve significantly over decision tree in terms of sensitivity and specificity, but ensemble classifier did. Overall, the best AUC, sensitivity, specificity, and accuracy were achieved by LogitBoost as an ensemble classifier (Figure 4)."

Discussion

- Under "Principal Findings," in the second paragraph, the sentence "*…(measured in pairwise resampled t tests, with P value adjusted to .008 using Bonferroni correction)*" has been replaced by "*…(measured in pairwise resampled t tests)*."
- Under "Principal Findings," in the last paragraph, the sentence "*…or those requiring higher cognitive abilities, such as assessing the propensity for changes and expressions of modality describing possibility, necessity, and certainty of health events and situations*" has been replaced by "*…or those requiring higher cognitive abilities, such as abstract terms denoting importance, significance, noticeability or markedness of health events and situations*."

Authors' Contributions

In the originally published paper, the following "Authors' Contributions" section was not included.

> *MJ and TH were responsible for overall research design; MJ was responsible for paper writing and revision, and YL was responsible for formal analysis and data curation.*

Multimedia Appendices

The information presented in the Multimedia Appendix 1 entitled "Variables in the logistic regression of health text understandability membership" has been updated. The originally published Multimedia Appendix 1 is in Multimedia Appendix 2.

Figures and Tables

Figures 1-4 have been replaced and can be viewed below. The originally published Figures 1-4 are in Multimedia Appendix 3. Tables 1-5 have been updated and can be viewed below. The originally published Tables 1-5 are in Multimedia Appendix 4.

Figure 1. Hyperparameter tuning (decision tree).

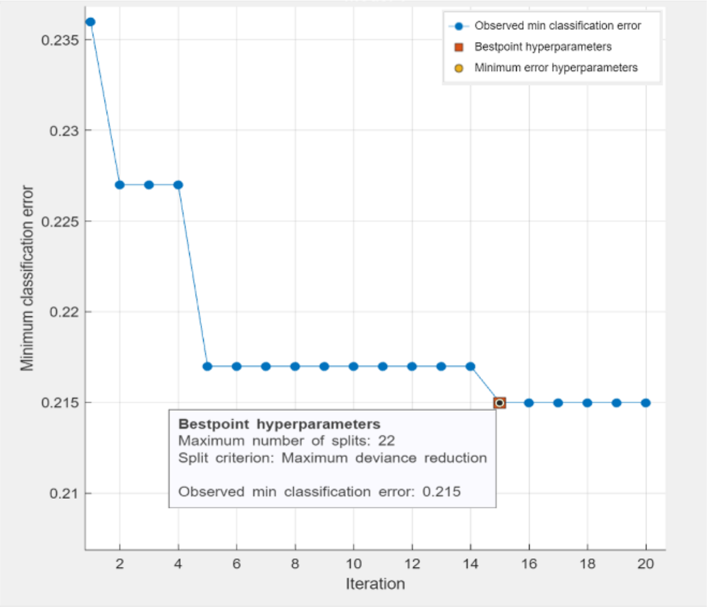


Figure 2. Hyperparameter tuning (ensemble classifier).

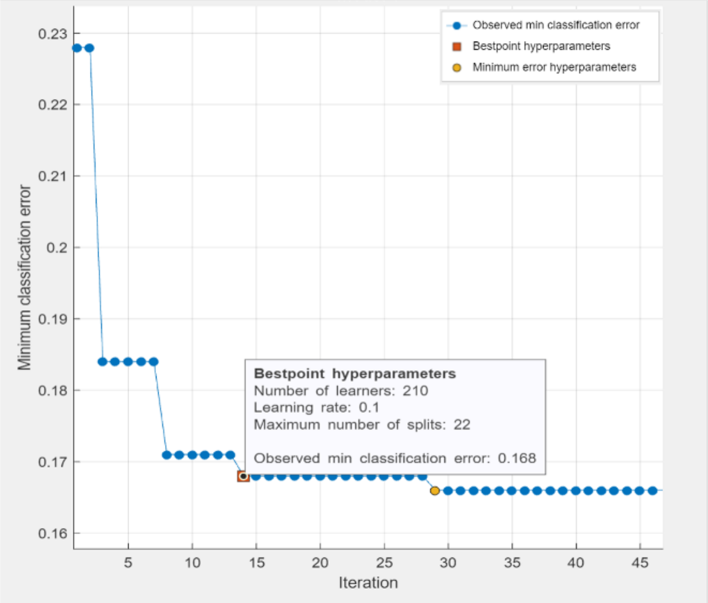


Figure 3. Hyperparameter tuning (support vector machine).

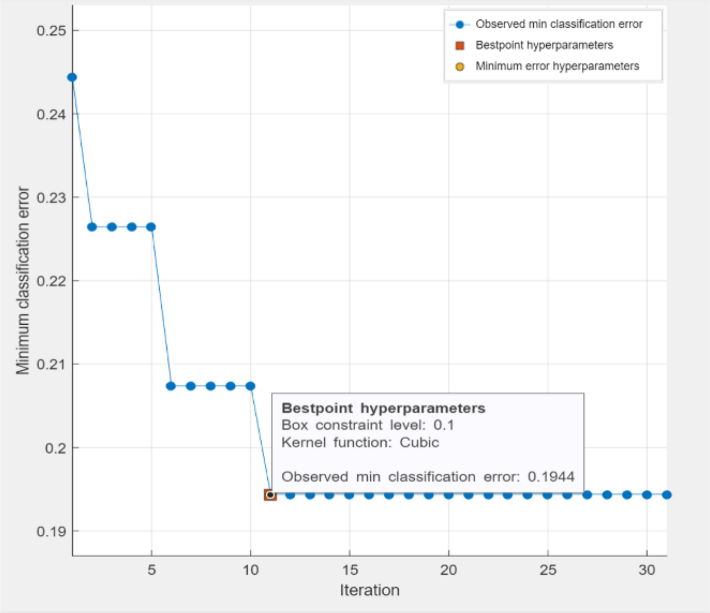


Figure 4. Mean receiver operating characteristic (ROC) curve for machine learning algorithms. LR: logistic regression; SVM: support vector machine; DT: decision tree; ENS: ensemble classifier (LogitBoost); ROC: receiver operating characteristic.

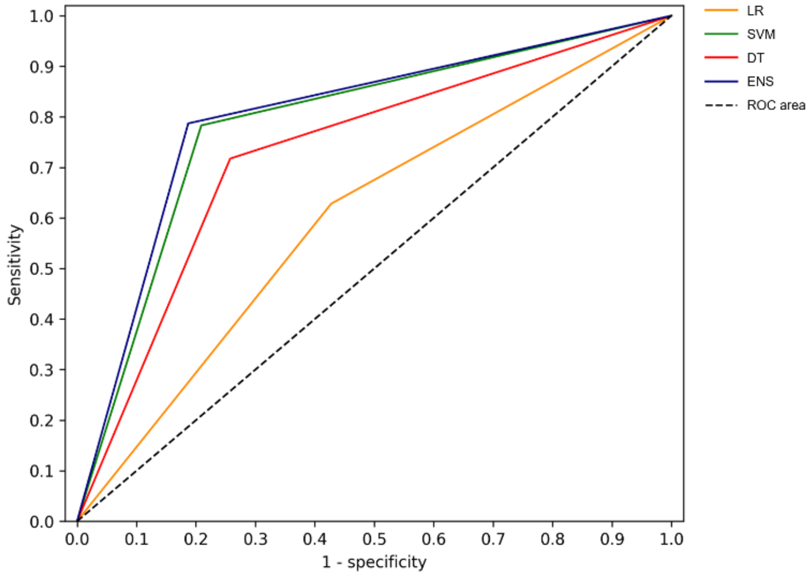


Table 1. Performance of the machine learning models using multidimensional semantic features as predictors.

Algorithm	AUC[a], mean (SD)	Sensitivity, mean (SD)	Specificity, mean (SD)	Accuracy, mean (SD)
LR[b]	0.614 (0.0554)	0.6282 (0.0597)	0.5724 (0.0733)	0.6010 (0.0523)
SVM[c]	0.848 (0.0172)	0.7830 (0.0368)	0.7910 (0.0420)	0.7860 (0.0153)
DT[d]	0.754 (0.0377)	0.7174 (0.0719)	0.7424 (0.0589)	0.732 (0.0317)
ENS[e]	0.858 (0.041)	0.787 (0.057)	0.813 (0.046)	0.802 (0.032)

[a] AUC: area under the operating characteristic curve.

[b] LR: logistic regression.

[c] SVM: support vector machine.

[d] DT: decision tree.

[e] ENS: ensemble classifier (LogitBoost).

Table 2. Pairwise corrected resampled *t* test of area under the curve differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.2340 (0.0669)	0.0299	–0.3171 to –0.1509	–7.817 (4)	.001
LR vs DT[c]	–0.1460 (0.0551)	0.0246	–0.2144 to –0.0777	–5.931 (4)	.004
LR vs ENS[d]	–0.2440 (0.0564)	0.0252	–0.3140 to –0.1740	–9.675 (4)	.001
SVM vs DT	0.0880 (0.0192)	0.0086	–0.0641 to 0.1119	10.230 (4)	.001
SVM vs ENS	–0.0100 (0.0374)	0.0167	–0.0565 to –0.0365	–0.598 (4)	.582
DT vs ENS	–0.0980 (0.0192)	0.0086	–0.1219 to –0.0741	–11.392 (4)	<.001

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 3. Pairwise corrected resampled *t* test of sensitivity differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.1548 (0.0303)	0.0135	–0.1924 to –0.1172	–11.429 (4)	<.001
LR vs DT[c]	–0.1002 (0.0720)	0.0322	–0.1896 to –0.0108	–3.111 (4)	.036
LR vs ENS[d]	–0.1588 (0.0945)	0.0423	–0.2761 to –0.0414	–3.756 (4)	.020
SVM vs DT	0.0546 (0.0697)	0.0312	–0.0319 to 0.1411	1.752 (4)	.155
SVM vs ENS	–0.0040 (0.0855)	0.0382	–0.1102 to –0.1022	–0.105 (4)	.922
DT vs ENS	–0.0586 (0.0371)	0.0166	–0.1046 to –0.0126	–3.535 (4)	.024

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 4. Pairwise corrected resampled *t* test of specificity differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.2186 (0.0968)	0.0433	–0.3389 to –0.0984	–5.047 (4)	.007
LR vs DT[c]	–0.1720 (0.0822)	0.0368	–0.2741 to –0.0699	–4.679 (4)	.009
LR vs ENS[d]	–0.2410 (0.0677)	0.0303	–0.3251 to –0.1569	–7.959 (4)	.001
SVM vs DT	0.0466 (0.1059)	0.0474	–0.0849 to 0.1781	0.984 (4)	.381
SVM vs ENS	–0.0224 (0.0918)	0.0411	–0.1364 to –0.0916	–0.545 (4)	.614
DT vs ENS	–0.0690 (0.0334)	0.0149	–0.1105 to –0.0275	–4.619 (4)	.010

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

Table 5. Pairwise corrected resampled *t* test of accuracy differences (using multidimensional semantic features as predictor variables).

Pairs	Mean difference (SD)	Standard error mean	95% CI	*t* test (*df*)	*P* value
LR[a] vs SVM[b]	–0.1850 (0.0507)	0.0227	–0.2480 to –0.1220	–8.152 (4)	.001
LR vs DT[c]	–0.1370 (0.0482)	0.0215	–0.1968 to –0.0771	–6.360 (4)	.003
LR vs ENS[d]	–0.2010 (0.0549)	0.0246	–0.2692 to –0.1328	–8.182 (4)	.001
SVM vs DT	0.0480 (0.0295)	0.0132	0.0114 to 0.0846	3.639 (4)	.022
SVM vs ENS	–0.0160 (0.0366)	0.0164	–0.0615 to 0.0295	–0.976 (4)	.384
DT vs ENS	–0.0640 (0.0148)	0.0066	–0.0823 to –0.0457	–9.704 (4)	.001

[a] LR: logistic regression.

[b] SVM: support vector machine.

[c] DT: decision tree.

[d] ENS: ensemble classifier (LogitBoost).

The authors confirm that the results and conclusions of the corrected data are consistent with those in the originally published version.

These corrections will appear in the online version of the paper on the JMIR website on September 21, 2021, together with the publication of this correction notice. Because this was made after submission to PubMed, PubMed Central, and other full-text repositories, the corrected article has also been resubmitted to those repositories.

Multimedia Appendix 1
Variables in the logistic regression of health text understandability membership.
[DOCX File , 34 KB - medinform_v9i9e33385_app1.docx]

Multimedia Appendix 2
Originally published Multimedia Appendix 1.
[DOCX File , 34 KB - medinform_v9i9e33385_app2.docx]

Multimedia Appendix 3
Originally published Figures 1-4.
[PDF File (Adobe PDF File), 542 KB - medinform_v9i9e33385_app3.pdf]

Multimedia Appendix 4
Originally published Tables 1-5.
[PDF File (Adobe PDF File), 1428 KB - medinform_v9i9e33385_app4.pdf]

Submitted 07.09.21; this is a non–peer-reviewed article;accepted 13.09.21; published 21.09.21.

Please cite as:
Ji M, Liu Y, Hao T
Correction: Predicting Health Material Accessibility: Development of Machine Learning Algorithms
JMIR Med Inform 2021;9(9):e33385
URL: https://medinform.jmir.org/2021/9/e33385
doi:10.2196/33385
PMID:34546943

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org