JMIR Medical Informatics

Impact Factor (2022): 3.2
Volume 8 (2020), Issue 5 ISSN 2291-9694 Editor in Chief: Christian Lovis, MD, MPH, FACMI

Contents

Original Papers

Reviews

Original Paper

Determining Factors Affecting Nurses' Acceptance of a Care Plan System Using a Modified Technology Acceptance Model 3: Structural Equation Model With Cross-Sectional Data

Kuei-Fang Ho[1,2], MSc; Pi-Chen Chang[2], RN, PhD; Maria Dyah Kurniasari[2,3], RN, MSc; Sri Susanty[2,4], RN, MSc; Min-Huey Chung[2,5], RN, PhD

[1]Department of Nursing, Ching Kuo Institute of Management and Health, Keelung, Taiwan

[2]School of Nursing, College of Nursing, Taipei Medical University, Taipei, Taiwan

[3]Department of Nursing, Faculty of Medicine and Health Science, Universitas Kristen Satya Wacana, Salatiga, Central Java, Indonesia

[4]Department of Nursing, Faculty of Medicine, University of Halu Oleo, Kendari, Southeast Sulawesi, Indonesia

[5]Department of Nursing, Shuang-Ho Hospital, Taipei Medical University, New Taipei City, Taiwan

Corresponding Author:
Min-Huey Chung, RN, PhD
School of Nursing
College of Nursing
Taipei Medical University
250 Wuxing Street
Taipei,
Taiwan
Phone: 886 2736 1661 ext 6317
Fax: 886 2377 2842
Email: minhuey300@tmu.edu.tw

Abstract

Background: Health information technology is used in nursing practice worldwide, and holistic patient care planning can serve as a guide for nursing practice to ensure quality in patient-centered care. However, few studies have thoroughly analyzed users' acceptance of care plan systems to establish individual plans.

Objective: Based on the technology acceptance model 3 (TAM3), a user technology acceptance model was established to explore what determines the acceptance of care plan systems by users in clinical settings.

Methods: Cross-sectional quantitative data were obtained from 222 nurses at eight hospitals affiliated with public organizations in Taiwan. Using the modified TAM3, the collected data were employed to analyze the determinants of user acceptance of a care plan system through structural equation modeling (SEM). We also employed moderated multiple regression analysis and partial least squares–SEM to test the moderating effects.

Results: We verified all significant effects from the use of a care plan system among bivariate patterns in the modified TAM3, except for moderating effects. Our results revealed that the determinants of perceived usefulness and perceived ease of use significantly influenced perceived usefulness and perceived ease of use, respectively. The results also indicated that nurses' perceptions of subjective norm (path coefficient=.25, $P<.001$), perceived ease of use (path coefficient=.32, $P<.001$), and perceived usefulness (path coefficient=.31, $P<.001$) had significantly positive effects on their behavioral intention to use the care plan system, accounting for 69% of the total explained variance.

Conclusions: By exploring nurses' acceptance of a care plan system, this study revealed relationships among the variables in TAM3. Our results confirm that the modified TAM3 is an innovative assessment instrument that can help managers understand nurses' acceptance of health information technology in nursing practice to enhance the adoption of health information technology.

(JMIR Med Inform 2020;8(5):e15686) doi:10.2196/15686

KEYWORDS

care plan system; technology acceptance model 3; behavioral intention

Introduction

Nurses' ability to develop detailed care plans considerably influences the quality of patient care [1]. Care plans are essential tools for promoting holistic care and are used to guide the practice of, communication about, and recording of the provided care in routine care settings [2,3]. Suitable individual care plans have been associated with correct medical observations and appropriate nursing diagnoses [4-6]. Therefore, it is reasonable to infer that such care plans lead to the appropriate implementation of care, accurate judgments of achieved patient goals, and clinically effective nursing interventions. In nursing environments, informatics has been used to improve data management and promote care planning [7]. With the help of information technology, a care plan system was developed to facilitate the planning, organization, coordination, and recording of the nursing process.

Several models have been proposed to examine the factors affecting individual reactions to information technology. For example, the user acceptance of technology model is the most popular model used to evaluate information systems [8]. The technology acceptance model (TAM) identifies why individuals adopt new technologies in various domains and is a popular topic of research in the information systems field. The original TAM contains two belief constructs, namely perceived usefulness (PU) and perceived ease of use (PEOU), which have been defined by Venkatesh and Davis [9] and Venkatesh [10], respectively (see Table 1). These constructs determine an individual's behavioral intention (BI) toward using information technology; PU has a stronger and more direct impact than does PEOU [9-11].

Venkatesh and Bala [11] developed a theoretical framework for TAM-related research by synthesizing prior research conducted on the TAM. This theoretical framework involves the social influence, systemic characteristics of determinants, individual differences, and facilitating conditions related to PU and PEOU. Social influence encompasses the social processes and mechanisms that shape individuals' perceptions of various aspects of a technology. Systemic characteristics refer to the identity of a system and can help individuals perceive the ease of use and usefulness of said system. Individual differences are personal characteristics or demographics that influence PEOU and PU. Finally, facilitating conditions refer to organizational infrastructure and support, which promote the adoption of a technology in a given context. Venkatesh and Bala [11] combined a theoretical model of the determinants of PEOU and PU with the original TAM and called this extended model TAM3. This model has since proven to be reliable and highly accurate for predicting and explaining user acceptance of various forms of information technology.

Theoretical processes such as social influence and cognitive instruments explain the relationship between PU and its determinants (ie, subjective norm [SN], image [IMG], job relevance [REL], and result demonstrability [RES]). SN and IMG are categorized as social influence processes, whereas REL and RES are system characteristics that reflect the effects of cognitive instrumental processes. Furthermore, according to the theoretical framework of TAM3, individual differences and facilitating conditions explain the determinants of PEOU through the anchoring and adjustment of human decision making. Anchoring involves four constructs, namely perception of external control (PEC), computer self-efficacy (CSE), computer anxiety (CANX), and computer playfulness (PLAY). These constructs reflect how individuals anchor the PEOU of a target system to their beliefs. The adjustment of perceived enjoyment (ENJ) and objective usability modifies individuals' PEOU of a target system. Objective usability is determined through the comparison of the amount of time spent by an expert with that spent by a novice to perform a task using the system [10]. The specific definitions of the determinants of PU and PEOU are provided in Table 1.

The variables of the original TAM have the power to predict nurses' technological acceptance of and intention to use information technology [12,13]. One study employed the original TAM to explore nurses' acceptance of a nursing information system for care planning. The researchers reported that PEOU and PU significantly influenced nurses' acceptance levels [14]. Zhang et al [15] conducted a study on the determinants of PU in the context of mobile homecare nursing to better understand the acceptance of a technology.

After reviewing the literature on user acceptance of a nursing information system, we noted that most studies were based on the original TAM only or theories regarding the determinants of PU. In addition to studying the relationships of REL and RES with PU, Zhang et al [15] observed that SN and IMG within an organization were significant antecedents of PU and that PU was the most influential factor in the adoption of mobile information technology by homecare nurses. In other words, to date, few studies have examined the determinants of PEOU or developed a combined model of the determinants of PEOU and PU in the context of nursing information system use. The care plan system in this study was developed by the North American Nursing Diagnosis Association on the basis of their classification system and was validated by our previous research [16]. This paper presents an empirical study on this care plan system that incorporated the modified TAM3 to explore the acceptance mechanism of a care plan system. The objectives of this study were to (1) identify the determinants of nurses' acceptance of a care plan system and (2) determine the influence of bivariate patterns in the modified TAM3 on the use of a care plan system.

Table 1. Definitions of constructs in the modified technology acceptance model 3.

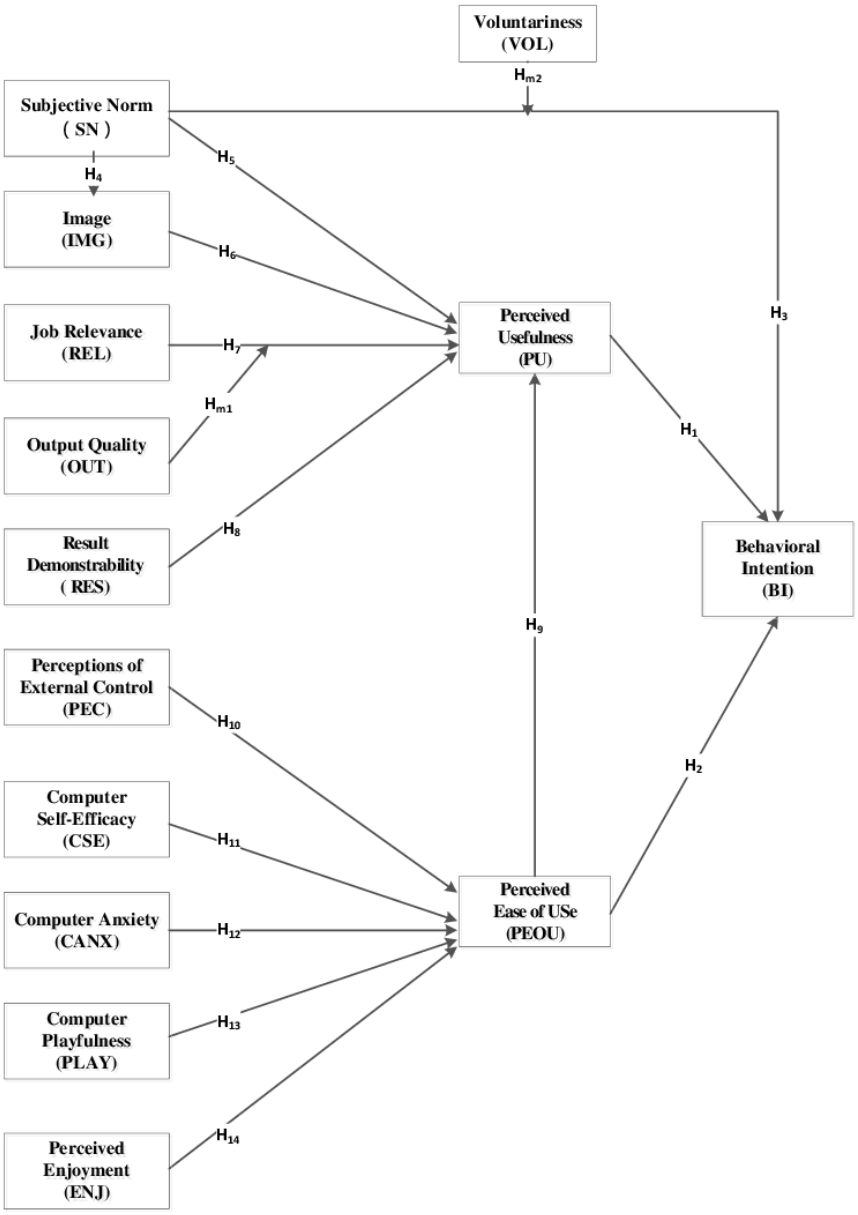
Construct	Definition
Perceived usefulness	The degree to which an individual believes that using a technology will enhance his or her job performance [9]
Social influence	
Subjective norm	An individual's perception of whether the people who are important to them think that they should use the target system [9]
Image	The degree to which an individual perceives that using a technology will enhance their image or status in their social circle [9,17]
Cognitive instruments	
Job relevance	One's perception of a technology as facilitative to their job [9]
Result demonstrability	An individual's perception of the tangible (observable and communicable) results from using the target system [9,17]
Perceived ease of use	The degree to which an individual believes that using a technology will be free of effort [10]
Anchoring	
Perception of external control	Individuals' perceptions regarding the availability of organizational responses to facilitate the use of a target system [11]
Computer self-efficacy	Individuals' beliefs regarding their abilities to use information technology [11]
Computer anxiety	An individual's degree of fear or apprehension when they use or consider using a target system [10,18]
Computer playfulness	The degree of perceived spontaneity in an individual's interaction with a technology [10]
Adjustment	
Perceived enjoyment	The performance-related consequences of using a target system and the degree to which using said system is perceived to be enjoyable [10]
Moderator	
Output quality	The strength of individuals' beliefs regarding how well a system enables the performance of a task with respect to said individuals' job goals [10]
Voluntariness	The rating range of voluntary use of a target system [10]

Methods

Theoretical Framework of the Technology Acceptance Model 3

This study proposed a modified version of TAM3, developed by Venkatesh and Bala [11], to express user acceptance of a care plan system. The study hypotheses are described as follows (Figure 1): (1) PU and PEOU have significant relationships with BI (H1 and H2, respectively); (2) the effects of SN on BI (H3), SN on IMG (H4), and SN and IMG on PU (H5 and H6, respectively) are related to social influence; (3) REL and RES represent the cognitive instrumental processes of PU (H7 and H8, respectively); (4) PEOU has a significant relationship with PU (H9); (5) the relationship of PEC, which refers to personnel beliefs, with PEOU (H10) is a facilitating condition; (6) the effects of CSE, PLAY, and CANX on PEOU (H11, H12, and H13, respectively) represent individual differences in terms of general beliefs about computers and computer use; and (7) ENJ can be adjusted to predict the PEOU of a system (H14). The degree to which the adjustment of objective usability determines the PEOU of a target system was not validated in this study.

Figure 1. Modified technology acceptance model 3 adopted in this study. Hm1: hypothesis moderator 1; Hm2: hypothesis moderator 2.



Venkatesh and Bala [11] included output quality (OUT) as a moderating variable (Table 1). REL on PU was stronger when OUT was higher. In addition, to distinguish voluntary use from mandatory use, the researchers included voluntariness (VOL) as a moderator (Table 1) of the relationship between SN and BI.

Study Design and Sample

This cross-sectional study was approved by the Medical Ethics Committee of the Tri-Service General Hospital (TSGHIRB No. B-104-13). The study period was from October 2015 to January 2016. All participants were registered nurses aged older than 20 years who had been using a care plan system for longer than 1 month. Using convenience sampling, 250 nurses were recruited from eight hospitals affiliated with public organizations in Taiwan. Data for this study were drawn from the same sample as that used in our previous study but were employed for different purposes and presented as a different type of data in this study. After informed consent was obtained from all participants, a structured questionnaire was employed for data collection.

Hair et al [19] proposed the estimation of the minimum sample size in partial least squares (PLS)–structural equation modeling (SEM) analysis with multiple regression models by applying Cohen [20] definitions; effect sizes of 0.02, 0.15, and 0.35 were considered small, medium, and large, respectively. To facilitate PLS-SEM analysis, the research sample size was calculated based on the recommendations of Hair et al [19]. The sample

size was calculated using the G*Power 3.0 software program (UCLA) with a power of .80, a medium effect size of 0.15, and alpha set at .05 for multiple regression of the maximum number of variables in a construct in our research framework, with five predictors used. A minimum sample size of 92 was necessary. Furthermore, a minimum sample size of 200 is often recommended for PLS-SEM [21,22]. Therefore, considering the 25% attrition rate, we recruited 250 nurses. The valid questionnaires completed by 222 nurses were used for data analysis, yielding a response rate of 88.80%.

Measures

Our questionnaire collected the demographic data of the nurses, and self-reported data were collected using the questionnaire about TAM3 designed by Venkatesh and Bala [11]. Following approval from the original author, 50 items in the modified TAM3 questionnaire composed the constructs investigated in our research model.

The modified TAM3 questionnaire consisted of the TAM constructs PU, PEOU, and BI; the determinants of PEOU (CSE, PEC, CANX, PLAY, and ENJ); the determinants of PU (SN, IMG, REL, and RES); and the moderators OUT and VOL. Except for the construct of CSE, items for all constructs were rated on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). The CSE items were measured on a 10-point Guttman scale ranging from 1 (strongly disagree) to 10 (strongly agree). The TAM3 questionnaire had high internal consistency reliability (Cronbach alpha ranging from .76 to .93) and high validity [11].

Data Analysis

Descriptive statistics were employed using SPSS Statistics version 20.0 (IBM Corp) to analyze sociodemographic variables and use characteristics of the care plan system. We estimated the measurement model, tested the structural model, and analyzed the relationships among all variables through PLS-SEM in SmartPLS version 3.0 (University of Hamburg).

Measurement Model Estimation

In accordance with the model evaluation criteria proposed by Hair et al [19], we assessed reliability, convergent validity, and discriminant validity. Internal consistency reliability was ensured if the composite reliability (CR) scores of all constructs and Cronbach alpha were higher than .70. Indicator reliability was ensured if all indicators' outer loadings were greater than .70. Convergent validity was confirmed if the average variance extracted (AVE) scores of all constructs were higher than .50. The square root of the AVE of each construct needed to be higher than the correlation between the latent variables, and all the indicators' outer loadings on their own constructs had to be higher than their cross-loadings with other constructs to satisfy the requirements of discriminant validity.

Moderating Effect Estimation

The PLS approach in SmartPLS and moderated multiple regression analysis in SPSS version 20.0 for Windows were applied to analyze and interpret interactions. We used SmartPLS version 3.0 to analyze the coefficients of interaction terms. The significance of a moderator was confirmed by t value (t >1.96) for all interaction effects (path coefficients). SPSS version 20.0 for Windows was used to calculate the model fit (R^2 without moderator), new model fit (R^2 with moderator), difference between these R^2 values, and significance of this difference for all endogenous latent variables.

Structural Model Analysis

To evaluate the multicollinearity of the structural model, two correlated variable correlation coefficients had to be <.85 [23]. A standardized root mean square residual (SRMR) lower than .10 indicated acceptable goodness of fit in the model [24]. The coefficient of determination values (R^2) representing weak, moderate, and substantial were .25, .50, and .75, respectively [19]. PLS-SEM with a bootstrapping procedure was employed to test the study hypotheses and analyze the path coefficients (significance level=5%).

Results

Participant Characteristics

The respondents reported their sociodemographic characteristics and use of the target information system. Of the 222 nurses, 95.5% (212/222) were women and 4.5% (10/222) were men. In total, 4 (1.8%) had a senior vocational school degree in nursing, 88 (39.6%) had an associate degree, and 130 (58.6%) had a bachelor's degree or higher. Most participants had more than 6 years of professional nursing experience (150/222, 67.6%). The use of health information technology for less than 6 years had the highest representation throughout the study sample (190/222, 85.6%). Most of the participants (141/222, 63.5%) did not feel under pressure when using a computer.

Measurement Model Results

As presented in Multimedia Appendix 1, for internal consistency reliability, all Cronbach alpha scores for the study variables were higher than .70, and CR scores ranged from .84 to .96, which were all acceptable. The outer loadings of all indicators were above .70, which implied satisfactory indicator reliability (see Multimedia Appendix 1). Multimedia Appendix 1 indicates that AVE scores for all variables were above .64. This result satisfied the requirement for convergent validity. To confirm the discriminant validity of constructs, we examined whether the square root of the AVE from each construct (see Multimedia Appendix 1) exceeded the correlation between the constructs in the research model. Moreover, as presented in Multimedia Appendix 2, we ensured that all indicators had outer loadings in relation to their own latent variables that were higher than their cross-loadings with other constructs. Therefore, we concluded that the measurement model satisfied the criteria for internal consistency, indicator, convergent, and discriminant validity.

Analysis of Moderating Effects

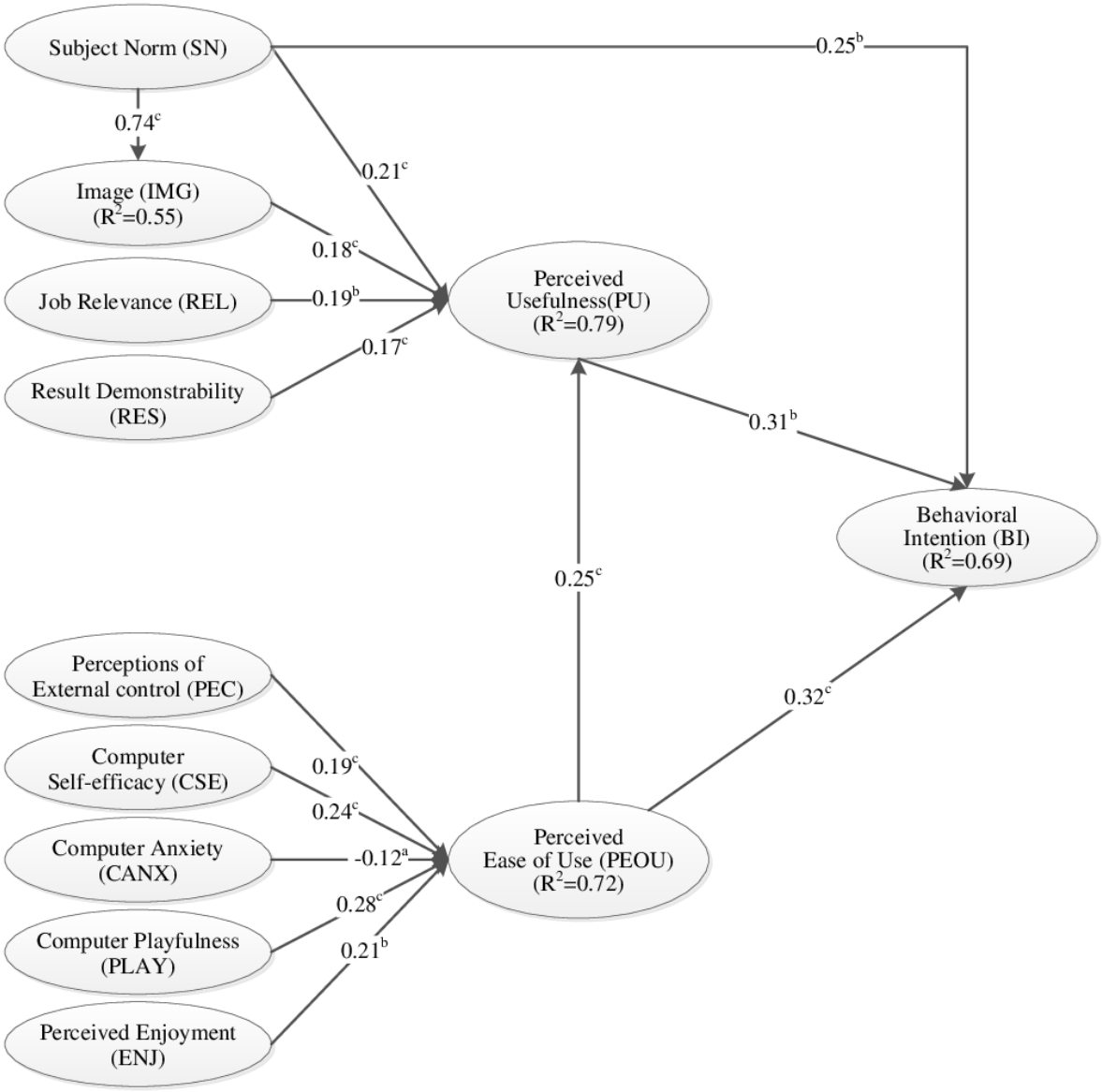
Moderated multiple regression analysis and PLS-SEM were employed to test the moderating effects. All test results are presented in Multimedia Appendix 3. The t values for all path coefficients were lower than 1.96, and differences among the R^2 values of all endogenous latent variables were minor and

nonsignificant. Therefore, VOL and OUT did not exert any moderating effects.

Structural Model Analysis and Hypothesis Testing

In this study, all bivariate correlations were lower than .85 (Multimedia Appendix 1). Therefore, multicollinearity was avoided. Multimedia Appendix 4 and Figure 2 present the explained variance of each construct. SN, IMG, REL, RES, and PEOU yielded approximately 79% of the variance for PU. The effects of CSE, PEC, CANX, PLAY, and ENJ on PEOU yielded approximately 72% of the total variance. The combination of SN, PEOU, and PU accounted for 69% of the variance observed for BI. This result indicated that the model explained high levels of variance.

Figure 2. Analysis path of the structural model. [a]*P*<.05. [b]*P*<.01. [c]*P*<.001. Note: No moderator variable was used in this model.



In our research model, the SRMR was .09, which indicated good model fit of the data. Therefore, the model was considered acceptable. The total indirect effect and total effect of all constructs on BI toward using the care plan system are presented in Multimedia Appendix 4, including the total indirect effects of SN (.11) and PEOU (.08) on BI. Combined with the direct effect, the total effects of SN (.36), PU (.31), and PEOU (.40) on BI were calculated.

A bootstrapping procedure was used to calculate the statistical significance of all path coefficients. Our researchers selected 5000 samples and recruited 222 nurses to estimate the path coefficients. As indicated by the PLS analysis results presented in Figure 2 and Multimedia Appendix 3 and Multimedia Appendix 4, all study hypotheses were supported by the data. The results revealed that SN, PEOU, and PU (path coefficients range=.25–.32) were all significant determinants of BI (R^2=.69). SN, IMG, REL, RES, and PEOU (path coefficients range=.17–.25) all had a significant effect on PU (R^2=.79). PEOU (R^2=.72) was significantly influenced by PEC, CSE,

PLAY, ENJ (path coefficients range=.19–.28) and CANX (path coefficient=−.12).

Discussion

Principal Findings

This study is the first to reveal the ability of TAM3 to comprehensively explore the determinants of BI for use of a care plan system. Our results indicated that the research model accounted for 69% of the variance in the care plan system, and all hypotheses supported the use of TAM3 except for the nonsignificant moderating effects of VOL and OUT. Few studies in nursing settings have explored user acceptance based on the determinants of PEOU and the combination of such determinants with those of PU. Our study empirically demonstrated that the determinants of PEOU influence PEOU and the determinants of PU and PEOU influence PU and SN, PEOU, with PU consequently predicting BI. Well-organized health information technology positively influences nurses' intentions to use a care plan system in professional settings [25]. This study provided an innovative methodology for evaluating and understanding nurses' acceptance of and need for a care plan system to implement well-organized health information technology and improve performance in nursing practice.

Using the modified TAM3, our research model explained 69% of the total variance, which was more than that explained by other TAM studies [12-15]. Our study results demonstrated that TAM3 is highly suitable for determining nurses' perceptions of using health information technology in nursing settings. Wu and Shen [26] indicated that PEOU, PU, and SN all had direct effects on health care professionals' BI. Moreover, in health care environments, PEOU and PU are key factors influencing the acceptance of health information technology by nursing personnel [12,15,27]. Using TAM3 with PU, PEOU, and SN to analyze users' BIs, we observed 69% variance for BI. In addition, SN, PEOU, and PU all had strong positive effects on BI, with path coefficients of .25 (P<.001), .32 (P<.001), and .31 (P<.01), respectively. In this study, the significant total effects of SN (path coefficient=.36), PU (path coefficient=.31), and PEOU (path coefficient=.40) on BI were also notable. Therefore, we assumed that the constructs of SN, PEOU, and PU are powerful predictors of nurses' BI to use a care plan system and contribute to the substantial explained variance of the modified TAM3. We suggest that implementing new health information technology in routine nursing care would improve related performance in nursing practice [25], broaden professional perspectives, and highlight preferences to enhance the ease of use of health information technology and improve key individual's opinions regarding the use of health information technology.

Using the modified TAM3, this study empirically verified the collected data and confirmed that the determinants of PEOU for measuring nurses' BI as well as all the determinants of PEOU had significant relationships with PEOU and explained 72% of the variance of PEOU. Moreover, we observed that PEOU had not only the most significant influence on BI to use the care plan system but also the strongest direct effect on BI. This result differed from those of previous studies [12,15,25]. In the pooled data for TAM3, PEC, CSE, PLAY (path coefficients range=.15–.33), and CANX (path coefficient=−.18) had a direct relationship with PEOU, and the total explained variance for PEOU was 52% [11]. The result of our research proved that PEC, CSE, PLAY, ENJ (path coefficients range=.19–.28), and CANX (path coefficient=−.12) significantly influence PEOU and jointly explain 72% of the variance in PEOU (R^2=.72). We posit that this research model with all the determinants of PEOU differed considerably from those used in previous studies, which adopted the modified TAM or the determinants of PU. That is, the TAM3 model provides a comprehensive set of PEOU determinants and an exhaustive explanation of the power of the PEOU of a care plan system. On the basis of our results, we recommend increasing individuals' BI to use computers to perform specific tasks, increase cognitive spontaneity related to computers, enhance enjoyment during the use of a target health information technology system, and reduce the level of fear in individuals' interactions with health information technology to promote nurses' PEOU toward the care plan system.

By comparing the direct effect of SN, IMG, REL, RES, and PEOU on PU in this study with the pooled data for TAM3 [11], we obtained SN, IMG, REL, RES, and PEOU values of .21/.04, .18/.24, .19/.03, .17/.26, and .25/.08, respectively. In this study, the determinants of PU explained 79% of the variance in PU (R^2=.79). As indicated in the pooled data of TAM3, PU is jointly predicted by the determinants of PU, with 67% of the total variance explained (R^2=.67) [11]. By contrast, other studies that adopted the modified model with the determinants of PU have predicted that PU accounts for 46% to 59% of the explained variance [15,28]. Our results were consistent with those of some previous studies [15,28], where nurses' PU of a care plan system was enhanced when they perceived that key individuals wanted them to use the health information technology in question. In addition, nurses' social status is enhanced when using said health information technology. Moreover, the significant relationships of REL (path coefficient=.19) and RES (path coefficient=.17) with PU in our study indicated that the nurses perceived health information technology as appropriate to their work. More tangibly, health information technology had a positive influence on the PU of the care plan system. When a user perceives that health information technology is useful, they also believe that it is easy to use [15]. Another key finding in our study was that PEOU had the most significant effect on PU. The determinants of PU in TAM3 are appropriate constructs for evaluating user belief regarding the usefulness of health information technology in nursing settings.

Venkatesh and Bala [11] argued that VOL and OUT are influential moderating variables in contexts where information technology is used. By contrast, our results revealed that the moderating variables VOL and OUT had no significant effects on the care plan system. Sun and Zhang [29] indicated that a weaker moderating effect elicits a stronger response from a more experienced user. The moderating effect of VOL weakens over time. Zhang and Cocosila [15] reported that the experience moderator did not influence homecare nurses' beliefs regarding the use of information technology. We assumed that all of our

participants had accumulated considerable experience of using a care plan system and that this led to the aforementioned nonsignificant moderating effects. The other reason for these effects may have been that our study adopted cross-sectional quantitative data to determine user acceptance, whereas TAM3 has generally been employed in longitudinal field studies. Therefore, the moderators had no significant effects.

Limitations and Recommendations

The first limitation of this study is that the cross-sectional data used were all collected at the same time. This could have yielded nonsignificant moderating effects. Moreover, our participants had already used the care plan system for more than 1 month. Therefore, this may have led to the moderating effects of VOL on the bivariables weakening with increasing experience. To avoid confusion in the results, the experience moderator was not measured in this study. We recommend that in the future, researchers explore the factors of user acceptance in the early stages of health information technology implementation and conduct longitudinal field studies.

Second, individual knowledge, attitude, and skill level with respect to nursing are crucial for designing patient-centered care plans and improving patient care quality [2,30]. Because the decision-making aspect of care planning varies from person to person and nursing students have insufficient nursing knowledge to design suitable care plans for patients, the measurement of objective usability—a comparison between the amounts of time spent by an expert and a novice to perform a task using the system—is conflicted. Therefore, we did not examine the objective usability variable in this study. To examine the relationship between objective usability and PEOU, future studies could employ a simple operating system, such as a patient physical data record system.

In health care, information technology developments adapt to changing needs [14]. To increase the use of information technology and improve its performance, we recommend that health care institutions adopt a model that measures nurses' perceptions of health information technology use to identify why the implementation of health information technology is accepted or rejected.

Conclusion

We applied TAM3 [11] to validate and measure determinants that affect the BI of nurses to use a care plan system. The critical determinants affecting nurses' acceptance of a care plan system were empirically examined. The results emphasize that SN, PEOU, and PU all predicted users' BI to use the care plan system, and the determinants of PU and PEOU significantly influenced PU and PEOU. This research contributes to the exploration of user acceptance and to a better understanding of care plan system use in routine nursing practice.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Cronbach alpha, composite reliability, outer loadings, average variance extracted, and Pearson correlation coefficients for the construct variables.
[DOCX File , 16 KB - medinform_v8i5e15686_app1.docx]

Multimedia Appendix 2
Outer loadings and cross-loadings of the study variables.
[DOCX File , 20 KB - medinform_v8i5e15686_app2.docx]

Multimedia Appendix 3
Path coefficients and results of the moderating effects analysis and research hypotheses.
[DOCX File , 25 KB - medinform_v8i5e15686_app3.docx]

Multimedia Appendix 4
Results of hypothesis testing, R2 calculation, and determining the total effect and total indirect effect for all variables with respect to behavioral intention to use.
[DOCX File , 16 KB - medinform_v8i5e15686_app4.docx]

References

1. Lee TT. Nursing diagnoses: factors affecting their use in charting standardized care plans. J Clin Nurs 2005 May;14(5):640-647. [doi: 10.1111/j.1365-2702.2004.00909.x] [Medline: 15840079]
2. Patiraki E, Katsaragakis S, Dreliozi A, Prezerakos P. Nursing care plans based on NANDA, nursing interventions classification, and nursing outcomes classification: the investigation of the effectiveness of an educational intervention in Greece. Int J Nurs Knowledge 2017 Apr;28(2):88-93 [FREE Full text] [doi: 10.1111/2047-3095.12120] [Medline: 26472136]
3. Ballantyne H. Developing nursing care plans. Nurs Stand 2016 Feb 24;30(26):51-57. [doi: 10.7748/ns.30.26.51.s48] [Medline: 26907149]

4. Gulanick M, Myers J. Nursing Care Plans—E-Book: Diagnoses, Interventions, and Outcomes. Philadelphia: Elsevier Health Sciences; 2016.
5. Griffiths J, Hutchings W. The wider implications of an audit of care plan documentation. J Clin Nurs 1999 Jan;8(1):57-65. [doi: 10.1046/j.1365-2702.1999.00217.x] [Medline: 10214170]
6. Keenan G, Yakel E, Tschannen D, Mandeville M. Documentation and the nurse care planning process. In: Hughes RG, editor. Patient Safety and Quality: An Evidence-Based Handbook for Nurses. Rockville: Agency for Healthcare Research and Quality; 2008.
7. Fuller C. Challenges in nursing informatics. URL: http://rn-journal.com/journal-of-nursing/challenges-in-nursing-informatics [accessed 2019-12-21]
8. Chuttur M. Working Papers on Information Systems. 2009. Overview of the technology acceptance model: origins, developments and future directions URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1289&context=sprouts_all [accessed 2020-03-13]
9. 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]
10. Venkatesh V. Determinants of perceived ease of use: integrating control, intrinsic motivation, and emotion into the technology acceptance model. Info Syst Res 2000 Dec;11(4):342-365. [doi: 10.1287/isre.11.4.342.11872]
11. 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]
12. Kuo K, Liu C, Ma C. An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. BMC Med Inform Decis Mak 2013 Aug 12;13(1):88. [doi: 10.1186/1472-6947-13-88]
13. Kowitlawakul Y. The technology acceptance model: predicting nurses' intention to use telemedicine technology (eICU). Comput Inform Nurs 2011 Jul;29(7):411-418. [doi: 10.1097/NCN.0b013e3181f9dd4a] [Medline: 20975536]
14. Dharmarajan B, Gangadharan K. Applying technology acceptance (TAM) model to determine the acceptance of nursing information system (NIS) for computer generated nursing care plan among nurses. Int J Comput Trends Technol 2013;4(8):2625-2629.
15. Zhang H, Cocosila M, Archer N. Factors of adoption of mobile information technology by homecare nurses: a technology acceptance model 2 approach. Comput Inform Nurs 2010;28(1):49-56. [doi: 10.1097/NCN.0b013e3181c0474a] [Medline: 19940621]
16. Ho K, Ho C, Chung M. Theoretical integration of user satisfaction and technology acceptance of the nursing process information system. PLoS One 2019;14(6):e0217622 [FREE Full text] [doi: 10.1371/journal.pone.0217622] [Medline: 31163076]
17. Moore GC, Benbasat I. Development of an instrument to measure the perceptions of adopting an information technology innovation. Info Syst Res 1991 Sep;2(3):192-222. [doi: 10.1287/isre.2.3.192]
18. 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 1995 Jan;3(2):231-247. [doi: 10.2190/7chy-5cm0-4d00-6jcg]
19. Hair J, Hult G, Ringle C, Sarstedt M. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). Thousand Oaks: Sage Publications; 2013.
20. Cohen J. A power primer. Psychol Bull 1992 Jul;112(1):155-159. [doi: 10.1037//0033-2909.112.1.155] [Medline: 19565683]
21. Marsh HW, Hau KT, Balla JR, Grayson D. Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivariate Behav Res 1998 Apr 01;33(2):181-220. [doi: 10.1207/s15327906mbr3302_1] [Medline: 26771883]
22. Hair J. Multivariate Data Analysis. Upper Saddle River: Pearson Prentice Hall; 2006.
23. Awang Z. Research Methodology and Data Analysis. Selangor: Penerbit Universiti Teknologi MARA Press; 2012.
24. Hu L, Bentler PM. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. Psychol Methods 1998;3(4):424-453. [doi: 10.1037/1082-989X.3.4.424]
25. Lu C, Hsiao J, Chen R. Factors determining nurse acceptance of hospital information systems. Comput Inform Nurs 2012 May;30(5):257-264. [doi: 10.1097/NCN.0b013e318224b4cf] [Medline: 22228251]
26. Wu J, Shen W, Lin L, Greenes RA, Bates DW. Testing the technology acceptance model for evaluating healthcare professionals' intention to use an adverse event reporting system. Int J Qual Health Care 2008 Apr;20(2):123-129. [doi: 10.1093/intqhc/mzm074] [Medline: 18222963]
27. Hsiao J, Wu W, Chen R. Factors of accepting pain management decision support systems by nurse anesthetists. BMC Med Inform Decis Mak 2013 Jan 29;13:16 [FREE Full text] [doi: 10.1186/1472-6947-13-16] [Medline: 23360305]
28. Yu P, Li H, Gagnon M. Health IT acceptance factors in long-term care facilities: a cross-sectional survey. Int J Med Inform 2009 Apr;78(4):219-229. [doi: 10.1016/j.ijmedinf.2008.07.006] [Medline: 18768345]
29. Sun H, Zhang P. The role of moderating factors in user technology acceptance. Int J Hum Comput Stud 2006 Feb;64(2):53-78. [doi: 10.1016/j.ijhcs.2005.04.013]
30. Tuinman A, de Greef MHG, Krijnen WP, Paans W, Roodbol PF. Accuracy of documentation in the nursing care plan in long-term institutional care. Geriatr Nurs 2017;38(6):578-583. [doi: 10.1016/j.gerinurse.2017.04.007] [Medline: 28552204]

Abbreviations

AVE: average variance extracted
BI: behavioral intention
CANX: computer anxiety
CR: composite reliability
CSE: computer self-efficacy
ENJ: perceived enjoyment
IMG: image
OUT: output quality
PEC: perception of external control
PEOU: perceived ease of use
PLAY: computer playfulness
PLS: partial least squares
PU: perceived usefulness
REL: job relevance
RES: result demonstrability
SEM: structural equation modeling
SN: subjective norm
SRMR: standardized root mean square residual
TAM: technology acceptance model
VOL: voluntariness

Edited by G Eysenbach; submitted 30.07.19; peer-reviewed by IC Lin, B Voshall, S Buchholz; comments to author 08.12.19; revised version received 02.02.20; accepted 24.02.20; published 05.05.20.

Please cite as:
Ho KF, Chang PC, Kurniasari MD, Susanty S, Chung MH
Determining Factors Affecting Nurses' Acceptance of a Care Plan System Using a Modified Technology Acceptance Model 3: Structural Equation Model With Cross-Sectional Data
JMIR Med Inform 2020;8(5):e15686
URL: https://medinform.jmir.org/2020/5/e15686
doi:10.2196/15686
PMID:32369033

Review

Factors Influencing the Adoption of Health Information Standards in Health Care Organizations: A Systematic Review Based on Best Fit Framework Synthesis

Lu Han[1], BSc; Jing Liu[1], MSc; Richard Evans[2], Dr med; Yang Song[1], BSc; Jingdong Ma[1], Dr med

[1]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

[2]College of Engineering, Design and Physical Sciences, Brunel University London, London, United Kingdom

Corresponding Author:
Jingdong Ma, Dr med
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science and Technology
No 13 Hangkong Road
Qiaokou District
Wuhan, 430030
China
Phone: 86 27 83692826
Email: jdma@hust.edu.cn

Abstract

Background: Since the early 1970s, health care provision has experienced rapid growth in the investment and adoption of health information technologies (HITs). However, the development and deployment of HITs has often been conducted in silos, at different organizational levels, within different regions, and in various health care settings; this has resulted in their infrastructures often being difficult to manage or integrate. Health information standards (ie, the set norms and requirements that underpin the deployment of HITs in health care settings) are expected to address these issues, yet their adoption remains to be frustratingly low among health care information technology vendors.

Objective: This study aimed to synthesize a comprehensive framework of factors that affect the adoption and deployment of health information standards by health care organizations.

Methods: First, electronic databases, including Web of Science, Scopus, and PubMed, were searched for relevant articles, with the results being exported to the EndNote reference management software. Second, study selection was conducted according to pre-established inclusion and exclusion criteria. Finally, a synthesized best fit framework was created, which integrated a thematic analysis of the included articles.

Results: In total, 35 records were incorporated into the synthesized framework, with 4 dimensions being identified: technology, organization, environment, and interorganizational relationships. The technology dimension included relative advantage, complexity, compatibility, trialability, observability, switching cost, standards uncertainty, and shared business process attributes. The organization dimension included organizational scale, organizational culture, staff resistance to change, staff training, top management support, and organizational readiness. The environment dimension included external pressure, external support, network externality, installed base, and information communication. Finally, the interorganizational relationships dimension included partner trust, partner dependence, relationship commitment, and partner power.

Conclusions: The synthesized framework presented in this paper extends the current understanding of the factors that influence the adoption of health information standards in health care organizations. It provides policy and decision makers with a greater awareness of factors that hinder or facilitate their adoption, enabling better judgement and development of adoption intervention strategies. Furthermore, suggestions for future research are provided.

(JMIR Med Inform 2020;8(5):e17334) doi:10.2196/17334

KEYWORDS

health information systems; health information interoperability; adoption; health care sector

Introduction

Background

During the last 50 years, the health care sector has experienced rapid technological growth, with the investment and adoption of health information technologies (HITs) showing promise to increase patient safety, reduce medical errors, improve efficiency, and reduce overall costs. However, health care systems are inherently complex, incorporating numerous interrelated and independent components [1]. A plethora of HITs exist across different levels of health care organizations [2]; however, underlying infrastructural issues have caused a multitude of integration and management issues [3]. This has resulted in many limitations, resulting in organizations not reaping the adoption benefits that were once promised, in particular, reduction in medical service costs [4]. For this reason, HITs should be adopted in a way that creates interoperability with other health care systems, enabling organizations to realize such benefits [5]. This can be resolved through the implementation of consensus standards [6]. The use of consensus standards is based on the idea of developing an agreed set of specifications or standards for data exchange that are not dependent on any proprietary software and are universally understood and accepted for data exchange [7].

Objective

Despite health information standards being seen as fundamental to the development of interoperable solutions [8,9], their adoption remains to be frustratingly low among information technology (IT) vendors and health organizations [10]. Prior studies have shown that the adoption of such standards in health care organizations is scarce [11-13]; however, there has been some exploration into the adoption of information standards not just in the health care sector. According to the results of these studies, different adoption factors may lead to difficulties for decision makers to explicitly understand, measure, and decrease inhibiting factors or enhance facilitating forces [14]. Hence, there is a need to synthesize those insights to provide decision makers with a holistic view of the adoption of health information standards. To achieve this goal and bridge the research divide, a comprehensive framework of factors that influence the adoption of health information standards is synthesized in this paper. The synthesized framework provides policy and decision makers with a more informed understanding of the factors that hinder or facilitate the adoption of health information standards, enabling better judgement and development of suitable strategies for adoption intervention. The following research questions (RQs) were proposed in this study:

- RQ1: What common factors have been included in previous studies that influence the adoption of health information standards by health care organizations?
- RQ2: Is there a framework that contains these factors more comprehensively from different dimensions?
- RQ3: If so, how will the adoption factors, included in the presented comprehensive framework, specifically affect the adoption of health information standards by health care organizations?

To answer these questions, this study aimed to identify and review existing articles on the adoption of information standards, extracting and summarizing their adoption factors to create a synthesized framework of the factors that affect the adoption of health information standards by health care organizations. A substantial number of stakeholders, including policy makers, citizens/patients, health care IT vendors, health care business owners, assessment bodies and regulators, clinicians and health care professionals, authorities and public administration departments, funders and health insurance companies, and academic departments, will find the presented framework beneficial in practice and when considering future research directions.

Methods

Study Design

A systematic review and framework synthesis were used as the methodological underpinning for our study. The systematic review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [15], whereas the *best fit* framework synthesis approach, proposed by Booth and Carroll [16], was adopted. The *best fit* approach is a relatively recent development, adapted from framework analysis, which involves systematically organizing data into an a priori conceptual framework. This study employed this approach for 2 reasons. First, the technology-organization-environment (TOE) framework, proposed by Tornatzky and Fleischer [17], seen as the most suitable framework for understanding technology adoption in organizational contexts, can be used as an a priori framework to integrate the factors that influence the adoption of health information standards. Second, although the approach is largely deductive (testing a framework), it also includes an inductive (thematic) analysis that is useful in understanding the phenomenon, especially the adoption of information standards from a health care perspective. Thus, this study will use the *best fit* approach to synthesize a comprehensive framework of factors affecting the adoption of health information standards by organizations based on the retrieved literature.

Search Strategy

This study comprehensively searched for all relevant literature in 3 electronic databases: Web of Science, Scopus, and PubMed. The search strategy employed is described in the following sections.

Web of Science

The Web of Science database was searched on July 25, 2019, and included 216 documents. The keywords used were as follows:

TS=(“information” OR “data”) AND TI=(“standards”) AND TI=(“adopt*” OR “accept*” OR “implement*”) AND TS=(“factors” OR “determinants” OR “barriers” OR ”facilitators”)

LANGUAGE=English

Scopus

The Scopus database was searched on July 25, 2019, and included 209 documents. The keywords used were as follows:

(TITLE-ABS-KEY(“information” OR “data”) AND TITLE(“standards”) AND TITLE(“adopt*” OR “accept*” OR “implement*”) AND TITLE-ABS-KEY(“factors” OR “determinants” OR “barriers” OR ”facilitators”))AND (LIMIT-TO(LANGUAGE, “English”))

PubMed

The PubMed database was searched on July 25, 2019, and included 36 documents. The keywords used were as follows:

((((information OR data)) AND standards[Title]) AND (adopt*[Title] OR accept*[Title] OR implement*[Title])) AND (factors OR determinants OR barriers OR facilitators). Filters: English

Inclusion Criteria

Studies were considered eligible (1) if they were related to the adoption of protocol, data sets, classification, coding, specification, terminology, identification, system framework, assessment, and other information or data standards; (2) if they involved research into the factors (including barriers and facilitators) that influence the adoption or implementation of standards; and (3) if they were based on relevant adoption theories, models, or frameworks, or if they involved the proposal of an adoption model or framework.

Exclusion Criteria

Studies considered ineligible for this research included those that (1) were not focused on the adoption of information or data standards; (2) did not involve factors that influenced standard adoption; or (3) did not involve relevant adoption theories, models, or frameworks.

Study Selection

In this study, search results were exported and indexed in EndNote X9.2, a reference management software. Once duplicates and patent documents were removed, LH screened the titles and abstracts of all remaining records for relevance. In the next step, the full-text articles of the retrieved results were examined by LH and JL for inclusion. Discrepancies were adjudicated by a senior researcher (JM).

Data Extraction and Synthesis

In this study, the *best fit* framework synthesis approach was followed, which integrates a thematic analysis to synthesize a comprehensive framework. The process consisted of the following stages:

1. *Familiarization with collected data.* On the basis of the understanding of the terminologies or terms used in the included studies, the factors influencing the adoption of information standards were initially extracted.
2. *Generation of initial codes.* According to definitions used in the identified studies, the extracted adoption factors were examined successively to make necessary mergers and trade-offs, generating a list of factors appropriate for health information standard adoption scenarios. The process included the following situations: (1) the factors with the same or similar meanings were combined into the same one and named with the most common term used in the literature; (2) the factors with different meanings were considered as juxtaposed dissimilar ones; and (3) if one factor was subordinate to another, the former was subsumed into the latter. For instance, *expected benefits* had the same meaning as *relative advantage*; these were combined into the same factor and named the latter. Similarly, *government support*, *vendor support*, and *partner support* were all related to *external support*, with the first three being subsumed into the last.
3. *Search for themes and define and name themes.* This stage consisted of 2 steps. First, the 3 dimensions of the prior framework (TOE) were used as initial themes for a deductive analysis, that is, based on the perceived commonality of the themes, the factors were analyzed and organized into 3 dimensions: technology, organization, and environment. The TOE framework explains that an organization’s decision to adopt technology can be jointly explained by 3 comprehensive dimensions, including technological, organizational, and environmental contexts. The technological context is essentially described by depicting the important attributes of the technology. The organizational context is depicted using descriptive measures concerning the organization (eg, scope, size, and managerial structure) and is influenced by formal and informal intraorganizational mechanisms for communication and control. The resources and innovativeness of the organization also play a role. The environmental context refers to the different attributes of the external environment in which an organization operates [18]. In the second step, apart from the 3 dimensions, another cluster of adoption factors, which could not be mapped against the TOE framework, was identified. The factors in this cluster were subsequently inductively analyzed, and a new dimension, titled interorganizational relationships, was generated. The interorganizational relationships are concerned with relationships between and among organizations, and it is a complex concept including many aspects, such as partner uncertainty, power, trust, and intermediary of relationship.
4. *Review themes.* This stage consisted of 2 levels. First, reviewing at the level of coded data. All adoption factors were reanalyzed within and across the dimensions to ensure consistency and independence. Second, reviewing at the level of themes. The dimensions were reviewed one final time to ensure they reflected the meaning of the adoption factors.

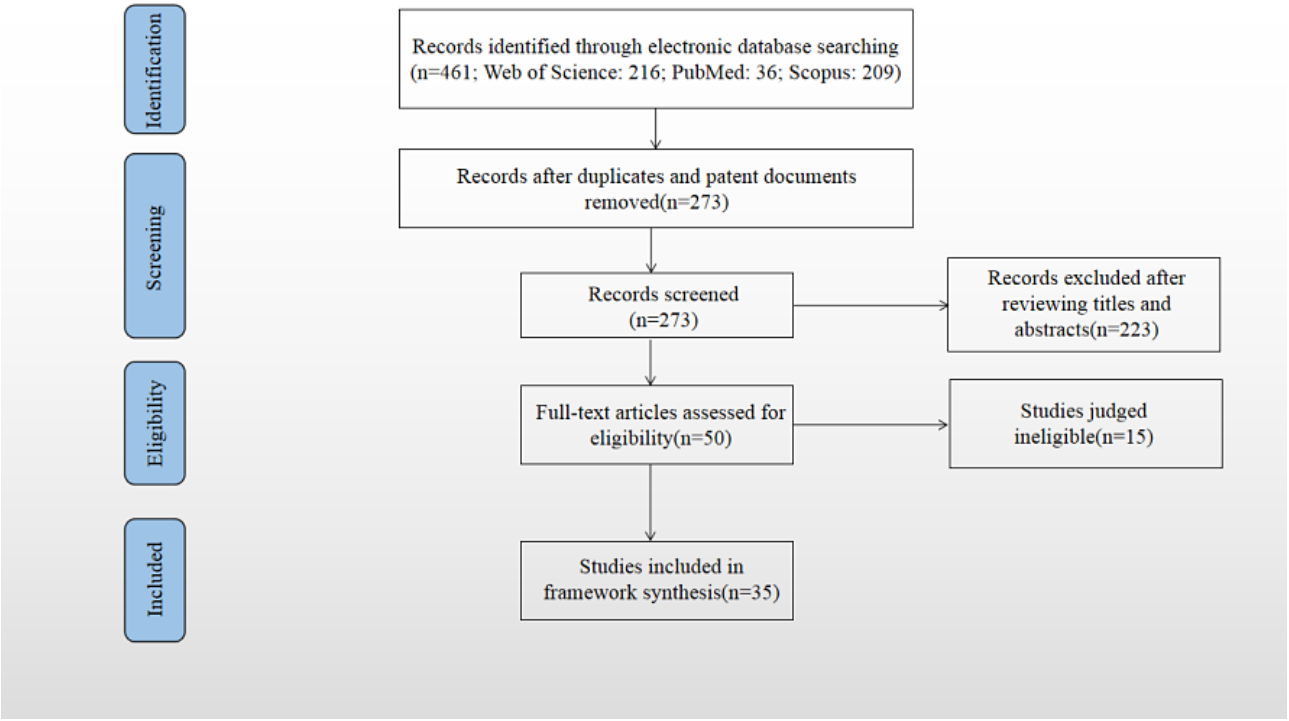
Ultimately, a comprehensive framework containing 4 dimensions (ie, technology, organization, environment, and interorganizational relationships) was synthesized. Throughout the synthesis, to ensure consistency in the classification of adoption factors, 3 researchers (LH, JL, and JM) discussed the factors to eliminate divergence.

Results

Search Results

In this study, 461 records were retrieved from the ISI Web of Science, Scopus, and PubMed databases. After removing 162 duplicate and 26 patent documents, the remaining 273 records were screened based on their titles and abstracts, according to the inclusion and exclusion criteria. As a result, 223 articles were deemed ineligible and were excluded. Then, after examining the full texts of the remaining 50 articles, 35 met the inclusion criteria and were included in the final review. Articles were excluded for the following reasons: 2 studies did not focus on the adoption of information or data standards; 2 studies did not involve factors that influence adoption; 2 studies were presented without relevant adoption theories, models, or frameworks; and 9 studies were not available in full. A flowchart summary of the literature search conducted is presented in the PRISMA diagram shown in Figure 1.

Figure 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.



Characteristics of Included Studies

The 35 articles included in this synthesis were mainly published from 2010 to 2018 (24/35, 68%). Among the included studies, 19 employed a quantitative design, 15 were qualitative, and 1 adopted a mixed methods approach. The quantitative studies mainly employed a questionnaire or survey, whereas the qualitative studies largely used interviews and focus group discussions. Eight studies were related to the adoption of information standards in the medical field, such as Health Level seven [13], health data standards [12,19,20], and data protection standards [21]. The remaining 27 focused on the IT field. For example, Internet Protocol version 6 [22-24], RosettaNet [25-27], and electronic data interchange [28-30]. Only 13 articles comprehensively considered the 3 dimensions of technology, organization, and environment [12,13,18,22,25,26,28,30-35], whereas one of them also included interorganizational determinants [28].

Results of Synthesis

This study took the adopting organization as the unit of analysis. On the basis of the *best fit* framework synthesis approach, the final synthesized framework included technology, organization, environment, and interorganizational relationships (Figure 2). The technology dimension incorporated relative advantage, complexity, compatibility, trialability, observability, switching cost, standards uncertainty, and shared business process attributes. The organization dimension included organizational scale, organizational culture, staff resistance to change, staff training, top management support, and organizational readiness. The environment dimension contained external pressure, external support, network externality, installed base, and information communication. Finally, the interorganizational relationships dimension included partner trust, partner dependence, relationship commitment, and partner power. These common factors identified in the included studies will have an impact on the adoption of health information standards by health care organizations. The specific impact of these factors will be detailed in the next section. The factors that influence the adoption of health information standards under the 4 dimensions are shown in Table 1 (for the definition of each factor, see Multimedia Appendix 1).

Figure 2. The synthesized framework.

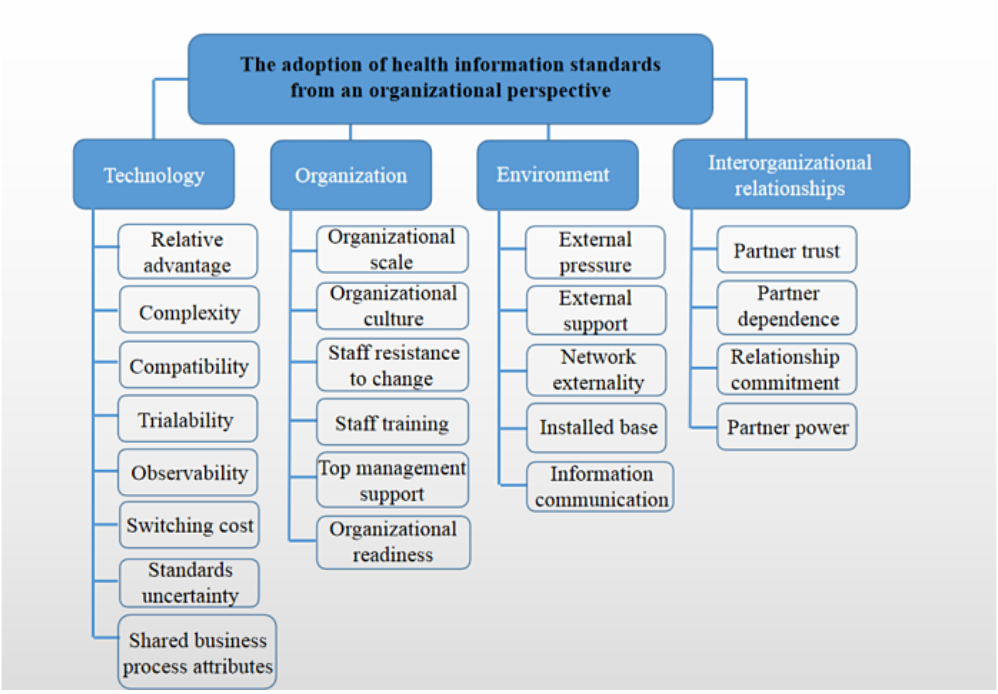


Table 1. Factors that influence the adoption of health information standards under the 4 dimensions.

Dimensions and factors	References
Technology	
Relative advantage	[22,24,26,28,30-41]
Complexity	[12,13,22,24,26,28,31-34,36,38,41,42]
Compatibility	[12,13,18,24-26,28,32-36,38,41,42]
Observability	[24,26,41,42]
Trialability	[24,26,34,36,41,43]
Switching cost	[12,23,34,44,45]
Standards uncertainty	[25]
Shared business process attributes	[35]
Organization	
Organizational scale	[13,26,28,31,32,38,41]
Organizational culture	[12,25,26,31,41,46]
Staff resistance to change	[12]
Staff training	[18,46]
Top management support	[18,22,26,28,32,35,37,38,40,42,46]
Organizational readiness	[26,28,30-32,34,35,38,40,41,43,44,47,48]
Environment	
External pressure	[18,22,23,27,30,37,43,48-50]
External support	[13,18,23,24,26,27,32,34,38,40,43,46,49]
Network externality	[12,24,28,34,44,45]
Installed base	[24,31,33]
Information communication	[24,50]
Interorganizational relationships	
Partner trust	[26-29,43]
Partner dependence	[28,29]
Relationship commitment	[26,28,29]
Partner power	[26,27]

Discussion

Principal Findings

As identified in the included studies, there exist various objects (standards), fields of inquiry, and methodological approaches when it comes to exploring factors that influence the adoption of health information standards; each study has its own specific object and approach (for details of included studies, see Multimedia Appendix 2). In all studies, the adoption factors were identified and selected according to relevant theories, models, or frameworks and the specific standard. The resulting differences may be partly because of the different characteristics of the adopted standards and their different requirements for the adoption environment. However, the factors that influence their adoption can be useful in better understanding the adoption of health information standards by organizations. This study sought to identify the contributing factors that influence the adoption of health information standards in the health care sector, providing a comprehensive synthesized framework. As previously mentioned, the adoption factors have been organized into 4 dimensions, as explained in the following sections.

Technology Dimension

The characteristics of innovation have been frequently studied in research relating to innovation adoption. Whether an innovation can be adopted by an individual, organization, or industry and its own characteristics and advantages, namely, its own technical factors, play a pivotal role. Therefore, the technical factors of the adopted standards are the primary consideration for the adoption of health information standards. The results of this study indicate that 19 studies used factors of technical characteristics for assessing the impact on adopting information standards [12,13,18,22-26,28,30-43,45]. In this study, relative advantage, complexity, compatibility, trialability, observability, switching cost, standards uncertainty, and shared business process attributes were included in the synthesized framework.

Roger identified 5 perceived attributes of an innovation that may determine the innovation's rate of adoption [51]. These attributes are relative advantage, complexity, compatibility, trialability, and observability, which are deemed useful for assessing the decision to adopt standards. The degree of relative advantage may be measured in economic terms, such as faster development, less maintenance, and cost saving [41]; these advantages could generate new markets, products, and services, which in turn create a competitive advantage for early adopters [24]. Thus, the greater the perceived relative advantage of the standard, the more rapid its rate of adoption will be [41]. The increased complexity of each standard increases the effort required to implement it and, therefore, reduces the number of potential adopters [24]. Thus, the more complex a standard, the less likely it is to be adopted by the organization. If the adopted standard is compatible with existing technologies or infrastructure, and consistent with past experiences of the organization, the organization will tend to upgrade to the new standard to gain a competitive advantage. Furthermore, if the adopted standard is of high trialability, the organization can reduce uncertainty and the risk of deploying the standard and obtain an increased perceived value through an initial pilot study, which will increase the organization's willingness to adopt the standard. Similarly, if the adopted standard has significant observable benefits and quantifiable advantages, it will reduce the perceived risk and make the organization more willing to adopt the standard. In summary, when the adoption of standards is perceived as having greater relative advantage, compatibility, trialability, observability, and less complexity, the organization will be more inclined to adopt standards [24,26,41].

The cost of switching between standards was observed as a negative factor to standard adoption by health care organizations [12,23,34]. Cost is typically associated with the unfamiliarity of the organization with existing resources and skills regarding the standards. For example, if there is a lack of experts who can deal with or lead the adoption, then it will cost large sums to consult relevant experts. As a result, a great deal of staff training, and a high degree of change management, will be required. Mapping issues from the old information infrastructure to the new standardized one will also be a real cost concern; thus, the organization will consider that it has already invested in their current infrastructure and will be reluctant to discard an amount of capital and equipment, as a result of the requirements of adopting the new standard [12].

Standards uncertainty represents the perception of whether the process specifications and associated technologies will be stable, over a certain period, and able to deliver the intended benefits [25]. As David and Greenstein [52] noted, a firm may not be willing to adopt a standard until it becomes a de facto standard in the industry. Thus, if decision makers perceive that the technology and processes required for standard adoption are not stable and are not going to change in the future, this will hinder the adoption of standards by organizations. Finally, as adopted standards are often based on business processes and information sharing between organizations, shared business process attributes, such as transaction volume needs, timeliness of exchange, effectiveness of communications, accuracy and integrity needs, and collaboration levels between participants, will influence the organization's decision on whether to adopt the standards or not [35].

Organization Dimension

Choosing whether to adopt standards or not is an organizational-level decision executed in an interorganizational context. There are various aspects of standard adoption that cannot be explained by technical factors alone [28]. Although the adoption of health information standards will promote better information sharing and connectivity within and between organizations, there are certain risks and uncertainties in adoption behavior because of past experiences; hence, organizational factors play a significant role in decision making. On the basis of our findings, 19 studies used organizational characteristic factors to assess the impact of adopting information standards [12,13,18,22,25,26,28,30-35,37,38,40-43,46-48]. In this study, organizational scale, organizational culture, staff resistance to change, staff training, top management support, and organizational readiness were included in the synthesized framework.

Organizational size makes a significant contribution to the adoption of standards [13,26,28,31,32,38,41]. According to some prior studies, large enterprises have several advantages over smaller ones. Large enterprises command considerable funds, talent, and research and development capacity, so they can realize the envisaged benefits quickly after adoption. On the contrary, other studies suggest that the bureaucracy of large enterprises is more complex and requires more time for decision making. Small- and medium-sized enterprises are effective and more conducive to adopting new technologies because of their efficient top-down introduction process; however, examination of the introduction effect may require further analysis to determine this conclusion.

Organizations that have a culture of innovation are more likely to experiment with standards at earlier stages [41]. Similarly, organizations should seek to strengthen internal knowledge management practices by constructing a learning organization, as knowledge management enables the knowledge of employees to evolve into the knowledge of the organization and teams. Organizations with rich knowledge of standard adoption are more likely to make decisions quicker and more effectively [31]. Furthermore, an organization's willingness to share information with its trading partners plays a key role in the success of standard adoption [26]. In short, organizations with a culture of innovation, learning, and information sharing are more likely to be early adopters of standards [12,25,26,31,41,46].

Alkraiji et al [12] established that employee reactions are a barrier to the adoption of health information standards because of the lack of understanding of the importance and benefits brought by standards. In addition, the staff's resistance to change also comes from their lack of relevant technical knowledge and ability. Sobol et al [53] indicated that the IT knowledge and capabilities of employees critically influence medical computerized system implementation; in other words, if the staff were more knowledgeable about standards, there would

be fewer advocator obstacles and less resistance against adoption.

Training is also deemed an important organizational mechanism that contributes to implementation success [54,55]. Employees must acquire new knowledge based on the understanding of the need for change to be able to overcome knowledge barriers and thus adopt new innovations effectively. Having an adequate training program is likely to increase employees' confidence and reduce resistance to standard adoption [56]. Moreover, training has been proven to enhance employee productivity and assist in utilizing the innovation to its full potential, which in turn can help organizations realize the full benefits derived from an innovation [57,58]. Therefore, the development of staff training effectively improves employees' relevant skills, capabilities, and knowledge of standard adoption, thus promoting the adoption process [18,46].

Previous studies have shown that top management support has a positive effect on technology adoption [13,59,60]. Senior managers can provide a long-term strategic vision, initiatives, and commitment to create a positive environment suitable for change [61]. Top management support can also enhance work satisfaction by modifying the rules and procedures that regulate and motivate employees' behavior to overcome the resistance to innovation implementation [62]. Young and Poon [63] asserted that top-level management support is essential in promoting interest and employees' satisfaction with the innovation. In the context of standard adoption, a high level of top management support means that top managers understand the benefits associated with the standard and demonstrate their commitment and political support. Therefore, top management support is expected to have a positive effect on standard adoption [18,22,26,28,32,35,38,42,46].

The success of innovation adoption is further dependent on an organization's preparation for the innovation. Organizational readiness, including technology readiness and resource readiness, can be used to measure an organization's capabilities for innovation adoption. Technology readiness refers to the level of sophistication of IT usage and management in an organization [35]. It includes top-level support from managers for related technologies [64], IT personnel, professional knowledge, skills, and experiences required for standard adoption [28,41]. Resource readiness measures whether an organization has enough resources to undertake the adoption [65]. It includes available financial resources to pay for installation costs, implementation of any subsequent enhancements, and ongoing expenses during usage [35], as well as other necessary resources, such as human resources, material resources, and information resources. If an organization has a high level of technology and resource readiness, it will have sufficient capacity to adopt standards, which will enable the organization to make decisions on standard adoption [26,28,30-32,34,35,38,41,43,48].

Environmental Dimension

All organizations exist in a certain social environment and will inevitably be affected by various external factors. When it comes to the adoption of health information standards, environment is a force that can encourage or impede an organization to adopt standards [28]; thus, environmental factors are also important factors that cannot be ignored. On the basis of the data extracted from the literature, 23 studies used environmental characteristic factors to assess the impact of adopting information standards [12,13,18,22-28,30-35,37,38,40,42,43,45,46,48-50]. In this study, external pressure, external support, network externality, installed base, and information communication were included in the synthesized framework.

An organization's decision to adopt standards is stimulated by pressures from various external sources, including the government [66-68], the industry in which it operates (ie, business partners and/or competitors) [69-71], and other sources, such as suppliers, customers, regulatory agencies, and professional associations [18]. Under the stimulation of these pressures, organizations may adopt relevant standards to seek sustainable development or actively strive for market competitiveness [18,22,23,27,30,43,48-50].

The level of external support is critical to the adoption of standards [13,18,23,24,26,27,32,34,38,43,46,49]. Morison [72] concluded that it is difficult for an organization to adopt a new standard without the intervention of an external agent in a position of power. Here, external support includes that from the government [30,73-75], which refers to governmental support for standard adoption through financial incentives, tax cuts, and pilot programs [49] and other forms of support that come from suppliers, external experts, and consultants [18,38,49], which will provide the organization with the necessary assistance and impetus to adopt standards.

Network externalities is one of the 2 main theories used within the stream of an economics perspective of standards and is related to the benefits created through the adoption of new standards by the potential community of adopters [12]. Positive network externalities provide support to the expectations of widespread adoption of a standard. Typically, the result is a reduction in cost because of the economies of scale and synergies created through increased opportunities of interactions among adopters [24]. As more organizations adopt the standard, barriers to adoption for others in the community are lowered [76,77]; thus, the network externalities have a positive influence on organizations to adopt standards [12,24,28,34].

In an internet environment that emphasizes interoperability, the large existing installed base and the resulting inertia (perception of switching costs and sunk costs) have a significant negative impact on the adoption of standards by organizations [24,31,33]. Farrell and Saloner [78] suggested that the current state of infrastructure, characterized by its installed base, the resulting inertia, and sunk costs in existing technology, can play an important role in determining the attractiveness of the environment for adoption. A well-established standard with a large installed base can create high drag and inertia, making the environment less attractive, thereby deterring organizations from adopting the new standard [24].

For an innovation to be adopted, information about it must be available to potential adopters [79,80]. The extent of information availability will depend on the level and nature of communication within the industry [81]. An environment with successful adoption cases and pioneering adopters can provide

favorable preconditions for information communication among organizations, thus raising awareness and encouraging innovation adoption [82]. Researchers view communication as vital to encourage the voluntary adoption of a new technology; this is because of a voluntary environment, where the lack of information might prompt other organizations to view the technology as risky, which fights against adoption [50]. Therefore, the effective communication of information relating to standards will play a positive role in propelling standard adoption by organizations [24,50].

Interorganizational Relationships Dimension

As the adoption of health information standards requires cooperation between two organizations, the relationship between an organization and its partner is salient [28]. In a technologically mature society, technology outsourcing becomes a prevalent method of satisfying an organization's technology needs. Technical issues become relatively insignificant compared with the interorganizational relationships in information standard adoption [83,84]. The major challenge is to build new electronic relationships [70,85]. Of the 35 included studies, 5 used factors of interorganizational relationships in assessing the impact on information standard adoption, which were grouped into a new dimension in this study [26-29,43]. Partner trust, partner dependence, relationship commitment, and partner power were included in the synthesized framework.

The trust between organizations lowers stress and improves adaptability [86]. In addition, information exchange is facilitated, and the effectiveness of joint problem solving is improved [87]. According to Shang et al [88], trust was an important factor in explaining interorganizational relationships. When business partners collaborated in their supply chains, an organization that trusted its partners was more likely to reach a consensus in terms of achievable benefits by the adoption of standards [27]. Thus, partner trust facilitates the adoption of standards by organizations [26-29,43].

Interdependence results from a relationship in which both organizations perceive mutual benefits from interaction [89] and in which any loss of autonomy will be equitably compensated through the expected gains. Both parties recognize that the advantages of interdependence provide benefits greater than those that either parties could attain by themselves [90]. Therefore, the interdependence will enable the partners to rely on each other and benefit from the adoption of standards based on a high degree of cooperation, which will facilitate the adoption of standards by organizations [28,29].

Another important antecedent for promoting standard adoption includes partner commitment to the trading relationship [29]. Commitment represents the willingness of trading partners to make efforts toward the relationship. Information standards requires a richer, more cooperative relationship [91]. The standard adopters working collectively with their trading partners can provide better service to customers (or suppliers), thereby increasing their market share [29]. Hence, if the partners can take coordinated actions, based on commitment to the relationship, it will be beneficial for both parties to reach a consensus on the adoption of standards [26,28,29].

It is possible that an organization may exert pressure on its trading partners to adopt standards based on partner power [26,27]. Partner power is defined as the capability of an organization to exert an influence on another organization to act in a prescribed manner [92]. Therefore, it is possible that in an interorganizational relationship, organizations with larger partner power can use compulsory or convincing power over their business partners in the adoption of standards [93].

The aforementioned factors that influence the adoption of information standards are extracted from the retrieved literature. The careful and comprehensive consideration and categorization of these factors yield a conceptual framework that can be used as a model for the adoption of health information standards, while remaining subject to adjustment and customization according to specific health information standards and the environment in which they are adopted. The adoption of health information standards can be illustrated in this conceptual framework against 4 dimensions: technological, organizational, environmental, and interorganizational relationships. Any consideration from a single perspective could be biased and fail to provide an accurate delineation of the phenomenon. However, it is worth noting that the synthesized conceptual framework was developed based on an extensive literature review related to information standard adoption and is currently in a preliminary stage. The relationships between the 4 dimensions contained in the framework and the relationships between the adoption factors and the adoption of health information standards by health care organizations can be examined through further empirical studies.

Limitations

This study has some limitations that should be acknowledged. First, because of the broad connotation of information standards, the search strategy employed in this study did not fully cover all concepts of information standards, which may lead to potential articles not being identified. Furthermore, because of resource constraints, the databases retrieved in this study were limited, which may result in other relevant studies not being retrieved. Second, this study excluded articles that did not involve relevant adoption theories, models, or frameworks and may have omitted some articles that solely proposed adoption factors. Finally, because of the overlap and intersection between the concepts of the adoption factors involved in the literature, there exists some subjectivity and bias in the concept definition and selection of factors and organizing the factors into corresponding dimensions in the synthesized framework.

In view of the above limitations, the synthesized framework may not include all possible adoption factors, which should be further improved and supplemented by research in the future. Nevertheless, this study has fully considered the factors that influence the adoption of health information standards, and the comprehensive framework provides references for future research and insights into the formulation and adoption of health information standards.

Conclusions

This study has comprehensively reviewed the factors that influence the adoption of information standards in the published

literature. A synthesized framework of integrated factors that influence the adoption of health information standards by organizations was extracted and presented.

This study delivers contributions at different levels. First, at the theoretical level, the synthesized framework has addressed knowledge gaps in the adoption of health information standards in health care organizations. Second, at the practice level, it will help guide policy and decision makers in better judging and developing suitable strategies for adoption interventions. For health care organizations, in particular, strategies for the adoption interventions include upgrading infrastructure and enriching technical resources and skills to better adapt to new standards; establishing an innovative culture, strengthening staff training, raising the attention of top managers, and increasing the investment of technology and resources to promote the implementation of new standards; heading on competitive pressure, leveraging external forces and information communication channels, and overcoming the industry inertia to actively respond to the adoption of new standards; and establishing trust and interdependency relationships among partners based on commitment and making reasonable use of partner power to create the industry fashion of standard adoption. Furthermore, it also provides directions for future research to enrich the factors that influence the adoption of relevant standards or health care technologies.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (HUST: 2019WKYXZX010).

Conflicts of Interest

None declared.

Multimedia Appendix 1
The definition of each factor.
[DOCX File , 16 KB - medinform_v8i5e17334_app1.docx]

Multimedia Appendix 2
Details of included studies.
[DOCX File , 40 KB - medinform_v8i5e17334_app2.docx]

References

1. Plsek PE, Greenhalgh T. Complexity science: The challenge of complexity in health care. Br Med J 2001 Sep 15;323(7313):625-628 [FREE Full text] [doi: 10.1136/bmj.323.7313.625] [Medline: 11557716]
2. Hakkinen H, Turunen P, Spil T. Information in Health Care Process - Evaluation Toolkit Development. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences. 2003 Presented at: HICSS'03; January 6-9, 2003; Big Island, HI, USA URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1174362&tag=1
3. Khoumbati K, Themistocleous M. Integrating the IT infrastructures in healthcare organisations: a proposition of influential factors. Electron J Gov 2006 Jan;4(1):27-36 [FREE Full text]
4. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med 2006 May 16;144(10):742-752. [doi: 10.7326/0003-4819-144-10-200605160-00125] [Medline: 16702590]
5. Park H, Hardiker N. Clinical terminologies: a solution for semantic interoperability. J Korean Soc Med Inform 2009 Mar;15(1):1-11. [doi: 10.4258/jksmi.2009.15.1.1]
6. Zhang Y, Xu Y, Shang L, Rao K. An investigation into health informatics and related standards in China. Int J Med Inform 2007 Aug;76(8):614-620. [doi: 10.1016/j.ijmedinf.2006.05.003] [Medline: 16793329]
7. Thomas JW. Loughborough University. 2018. The Adoption and Diffusion of Data-exchange Standards URL: https://hdl.handle.net/2134/34998 [accessed 2019-08-20]
8. Kahn MG, Bailey LC, Forrest CB, Padula MA, Hirschfeld S. Building a common pediatric research terminology for accelerating child health research. Pediatrics 2014 Mar;133(3):516-525 [FREE Full text] [doi: 10.1542/peds.2013-1504] [Medline: 24534404]
9. Berler A, Tagaris A, Angelidis P, Koutsouris D. A roadmap towards healthcare information systems interoperability in Greece. J Telecommun Inf Technol 2006;2006(2):59-73 [FREE Full text] [doi: 10.1117/12.2191918]
10. Hammond WE. The making and adoption of health data standards. Health Aff (Millwood) 2005;24(5):1205-1213. [doi: 10.1377/hlthaff.24.5.1205] [Medline: 16162564]
11. Olsen J, Baisch MJ. An integrative review of information systems and terminologies used in local health departments. J Am Med Inform Assoc 2014 Feb;21(e1):e20-e27 [FREE Full text] [doi: 10.1136/amiajnl-2013-001714] [Medline: 24036156]
12. Alkraiji AI, Jackson T, Murray I. Factors impacting the adoption decision of health data standards in tertiary healthcare organisations in Saudi Arabia. J Ent Inf Manag 2016 Sep;29(5):650-676. [doi: 10.1108/jeim-11-2014-0111]

13. Lin C, Lin I, Roan J, Yeh J. Critical factors influencing hospitals' adoption of HL7 version 2 standards: an empirical investigation. J Med Syst 2012 Jun;36(3):1183-1192. [doi: 10.1007/s10916-010-9580-2] [Medline: 20827568]
14. Li J, Talaei-Khoei A, Seale H, Ray P, Macintyre CR. Health care provider adoption of eHealth: systematic literature review. Interact J Med Res 2013 Apr 16;2(1):e7 [FREE Full text] [doi: 10.2196/ijmr.2468] [Medline: 23608679]
15. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64. [doi: 10.7326/0003-4819-151-4-200908180-00135] [Medline: 19622511]
16. Booth A, Carroll C. How to build up the actionable knowledge base: the role of 'best fit' framework synthesis for studies of improvement in healthcare. BMJ Qual Saf 2015 Nov;24(11):700-708 [FREE Full text] [doi: 10.1136/bmjqs-2014-003642] [Medline: 26306609]
17. Tornatzky LG, Fleischer M. Processes of Technological Innovation. Lexington, MA: Lexington Books; 1991.
18. Vatanasakdakul S, Aoun C, Chen Y. Chasing success: an empirical model for IT governance frameworks adoption in Australia. Sci Technol Soc 2017;22(2):182-211. [doi: 10.1177/0971721817702278]
19. Alkraiji A, Jackson T, Murray I. Barriers to the widespread adoption of health data standards: an exploratory qualitative study in tertiary healthcare organizations in Saudi Arabia. J Med Syst 2013 Apr;37(2):9895. [doi: 10.1007/s10916-012-9895-2] [Medline: 23321966]
20. Alkraiji A, Jackson T, Murray I. Health data standards and adoption process: Preliminary findings of a qualitative study in Saudi Arabia. Campus Wide Inf Syst 2011;28(5):345-359. [doi: 10.1108/10650741111181616]
21. Foth M, Schusterschitz C, Flatscher‐Thöni M. Technology acceptance as an influencing factor of hospital employees' compliance with data‐protection standards in Germany. J Public Health 2012;20(3):253-268. [doi: 10.1007/s10389-011-0456-9]
22. Wang X, Zander S. Extending the model of internet standards adoption: A cross-country comparison of IPv6 adoption. Inf Manag 2018 Jun;55(4):450-460. [doi: 10.1016/j.im.2017.10.005]
23. Hovav A, Hemmert M, Kim YJ. Determinants of internet standards adoption: The case of South Korea. Res Policy 2011 Mar;40(2):253-262. [doi: 10.1016/j.respol.2010.09.016]
24. Hovav A, Patnayakuni R, Schuff D. A model of internet standards adoption: the case of IPv6. Inform Syst J 2004;14(3):265-294. [doi: 10.1111/j.1365-2575.2004.00170.x]
25. Venkatesh V, Bala H. Adoption and impacts of interorganizational business process standards: Role of partnering synergy. Inf Syst Res 2012 Dec;23(4):1131-1157. [doi: 10.1287/isre.1110.0404]
26. Chan FT, Chong AY. A SEM–neural network approach for understanding determinants of interorganizational system standard adoption and performances. Decis Sup Syst 2012 Dec;54(1):621-630. [doi: 10.1016/j.dss.2012.08.009]
27. Yee‐Loong Chong A, Ooi K. Adoption of interorganizational system standards in supply chains: An empirical analysis of RosettaNet standards. Indus Manag Data Syst 2008;108(4):529-547. [doi: 10.1108/02635570810868371]
28. Huang Z, Janz BD, Frolick MN. A comprehensive examination of internet-EDI adoption. Inf Syst Manag 2008;25(3):273-286. [doi: 10.1080/10580530802151228]
29. Lee S, Lim GG. The impact of partnership attributes on EDI implementation success. Inf Manag 2003 Dec;41(2):135-148. [doi: 10.1016/s0378-7206(03)00043-0]
30. Kuan KK, Chau PY. A perception-based model for EDI adoption in small businesses using a technology–organization–environment framework. Inf Manag 2001 Oct;38(8):507-521. [doi: 10.1016/s0378-7206(01)00073-8]
31. Hu C. Main Factors Affecting the Adoption and Diffusion of Web Service Technology Standards. In: Proceedings of the International Conference on Information and Management Engineering. 2011 Presented at: ICCIC'11; September 17-18, 2011; Wuhan, China p. 81-87. [doi: 10.1007/978-3-642-24091-1_12]
32. Burbano A, Rardin R, Pohl E. Exploring the Factors Affecting the Identification Standards Adoption Process in the US Healthcare Supply Chain. In: 2011 Proceedings of PICMET'11: Technology Management in the Energy Smart World. 2011 Presented at: PICMET'11; July 31- August 4, 2011; Portland, OR, USA URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6017680
33. Kelly D, Feller J, Finnegan P. Complex Network-Based Information Systems (CNIS) Standards: Toward an Adoption Model. In: Proceedings of the IFIP International Working Conference on the Transfer and Diffusion of Information Technology for Organizational Resilience. 2006 Presented at: TDIT'06; June 7-10, 2006; Galway, Ireland p. 3-20. [doi: 10.1007/0-387-34410-1_1]
34. Wapakabulo J, Dawson R, Probets S, King T. A Step Towards the Adoption of Data-exchange Standards: A UK Defence Community Case Study. In: Proceedings of the 4th Conference on Standardization and Innovation in Information Technology. 2005 Presented at: SIIT'05; September 21-23, 2005; Geneva, Switzerland. [doi: 10.1109/siit.2005.1563812]
35. Nelson ML, Shaw MJ. IDEALS @ Illinois. 2003. The Adoption and Diffusion of Interorganizational System Standards and Process Innovations URL: http://hdl.handle.net/2142/84534 [accessed 2019-08-20]
36. Ramoni RB, Etolue J, Tokede O, McClellan L, Simmons K, Yansane A, et al. Adoption of dental innovations: The case of a standardized dental diagnostic terminology. J Am Dent Assoc 2017 May;148(5):319-327 [FREE Full text] [doi: 10.1016/j.adaj.2017.01.024] [Medline: 28364948]

37. Mueller T, Dittes S, Ahlemann F, Urbach N, Smolnik S. Because Everybody is Different: Towards Understanding the Acceptance of Organizational IT Standards. In: Proceedings of the 2015 48th Hawaii International Conference on System Sciences. 2015 Presented at: HICSS'15; January 5-8, 2015; Kauai, HI, USA. [doi: 10.1109/hicss.2015.487]
38. MacLennan E, van Belle J. Factors affecting the organizational adoption of service-oriented architecture (SOA). Inf Syst Bus Manag 2014;12(1):71-100. [doi: 10.1007/s10257-012-0212-x]
39. Singh RM, Dahlin K. Merit, Acceptance or Access: Opposing Forces to Adoption of a New Standard. In: Proceedings of the 5th International Conference on Standardization and Innovation in Information Technology. 2007 Presented at: SIIT'07; October 17-19, 2007; Calgary, AB, Canada. [doi: 10.1109/siit.2007.4629316]
40. Ng CS, Hsu PY, Tsai WH. Salient Factors for Maintenance Standard Adoption in Enterprise Resource Planning Context: An Exploratory Study. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences. 2006 Presented at: HICSS'06; January 4-7, 2006; Kauia, HI, USA, USA. [doi: 10.1109/hicss.2006.426]
41. Chen M. Factors affecting the adoption and diffusion of XML and Web services standards for E-business systems. Int J Hum Comput Stud 2003 Mar;58(3):259-279. [doi: 10.1016/s1071-5819(02)00140-4]
42. Velleman EM, Nahuis I, van der Geest T. Factors explaining adoption and implementation processes for web accessibility standards within eGovernment systems and organizations. Univ Access Inf Soc 2017;16(1):173-190. [doi: 10.1007/s10209-015-0449-5]
43. Henning F. Adoption of Interoperability Standards in Government Information Networks: An Initial Framework of Influence Factors. In: Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance. 2013 Presented at: ICEGOV'13; October 22 - 25, 2013; Seoul, Korea p. 264-267. [doi: 10.1145/2591888.2591936]
44. Gong N. Barriers to adopting interoperability standards for cyber threat intelligence sharing: an exploratory study. In: Arai K, Kapoor S, Bhatia R, editors. Intelligent Computing. Cham: Springer; 2018:666-684.
45. Techatassanasoontorn AA, Suo S. Influences on standards adoption in de facto standardization. Inf Technol Manag 2011;12(4):357-385. [doi: 10.1007/s10799-011-0089-2]
46. Ghahramani A. Factors that influence the maintenance and improvement of OHSAS 18001 in adopting companies: A qualitative study. J Clean Prod 2016 Nov;137:283-290. [doi: 10.1016/j.jclepro.2016.07.087]
47. Buyle R, van Compernolle M, Vlassenroot E, Vanlishout Z, Mechant P, Mannens E. 'Technology readiness and acceptance model' as a predictor for the use intention of data standards in smart cities. Media Commun 2018 Dec;6(4):127-139. [doi: 10.17645/mac.v6i4.1679]
48. Veit D, Parasie NP. Common Data Exchange Standards: Determinants for Adoption at the Municipal Level. In: Proceedings of the 2010 Americas Conference on Information Systems. 2010 Presented at: AMCIS'10; August 12-15, 2010; Lima, Peru URL: https://pdfs.semanticscholar.org/c600/9383c722b640f1249885f2e70a1b1459a9bc.pdf
49. Lucho S, Melendez K, Dávila A. Analysis of environmental factors in the adoption of ISO/IEC 29110. Multiple case study. In: Mejia J, Muñoz M, Rocha Á, Quiñonez Y, Calvo-Manzano J, editors. Trends and Applications in Software Engineering. Cham: Springer; Oct 2017:82-93.
50. Azam S. Perceived environmental factors and the intention to adopt a standard business reporting facility: A survey of Australian corporate CFOs. Asian Acad Manag J Account Finance 2014;10(2):147-173 [FREE Full text]
51. Rogers EM. Diffusion of Innovations. Fourth Edition. New York: Free Press; 1995.
52. David PA, Greenstein S. The economics of compatibility standards: an introduction to recent research. Econ Innov New Technol 1990;1(1-2):3-41. [doi: 10.1080/10438599000000002]
53. Sobol MG, Alverson M, Lei D. Barriers to the adoption of computerized technology in health care systems. Top Health Inf Manage 1999 May;19(4):1-19. [Medline: 10387652]
54. Dezdar S, Sulaiman A. Successful enterprise resource planning implementation: taxonomy of critical factors. Indus Manag Data Syst 2009 Sep;109(8):1037-1052. [doi: 10.1108/02635570910991283]
55. Hwang MI, Lin CT, Lin JW. Organizational Factors for Successful Implementation of Information Systems: Disentangling the Effect of Top Management Support and Training. In: Proceedings of the Southern Association for Information Systems Conference. 2012 Presented at: Proceedings of the Southern Association for Information Systems Conference; March 23-24, 2012; Atlanta, GA, USA p. 111-115 URL: https://tinyurl.com/y8su9emh
56. Dezdar S, Ainin S. The influence of organizational factors on successful ERP implementation. Manag Decis 2011;49(6):911-926. [doi: 10.1108/00251741111143603]
57. Liu PL. Empirical study on influence of critical success factors on ERP knowledge management on management performance in high-tech industries in Taiwan. Expert Syst Appl 2011;38(8):10696-10704. [doi: 10.1016/j.eswa.2011.02.045]
58. Tharenou P, Saks AM, Moore C. A review and critique of research on training and organizational-level outcomes. Hum Res Manag Rev 2007 Sep;17(3):251-273. [doi: 10.1016/j.hrmr.2007.07.004]
59. Vykoukal J. Grid Technology as Green IT Strategy? Empirical Results from the Financial Services Industry. In: Proceedings of the 18th European Conference on Information Systems. 2010 Presented at: ECIS'10; June 7-9, 2010; Pretoria, South Africa URL: https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1040&context=ecis2010
60. Lai H, Lin I, Tseng L. High-level managers' considerations for RFID adoption in hospitals: an empirical study in Taiwan. J Med Syst 2014 Feb;38(2):3. [doi: 10.1007/s10916-013-0003-z] [Medline: 24445396]

61. Yoon TE, George JF. Why aren't organizations adopting virtual worlds? Comput Hum Behav 2013 May;29(3):772-790. [doi: 10.1016/j.chb.2012.12.003]
62. Purvis RL, Sambamurthy V, Zmud RW. The assimilation of knowledge platforms in organizations: An empirical investigation. Organ Sci 2001 Apr;12(2):117-135. [doi: 10.1287/orsc.12.2.117.10115]
63. Young R, Poon S. Top management support—almost always necessary and sometimes sufficient for success: Findings from a fuzzy set analysis. Int J Proj Manag 2013 Oct;31(7):943-957. [doi: 10.1016/j.ijproman.2012.11.013]
64. Lee JN, Kim YG. Effect of partnership quality on IS outsourcing success: conceptual framework and empirical validation. J Manag Inf Syst 1999;15(4):29-61 [FREE Full text] [doi: 10.1080/07421222.1999.11518221]
65. Swatman PM, Swatman PA. EDI system integration: A definition and literature survey. Inform Soc 1992;8(3):169-205. [doi: 10.1080/01972243.1992.9960119]
66. Teo HH, Tan B, Wei KK. Innovation Diffusion Theory as a Predictor of Adoption Intention for Financial EDI. In: Proceedings of the 16th International Conference on Information Systems. 1995 Presented at: ICIS'95; December 10-13, 1995; Amsterdam, Netherlands p. 155-165 URL: https://pdfs.semanticscholar.org/cdd5/8f260875e98175cb690fa232d0e847a0c2c0.pdf
67. Lippert SK, Govindarajulu C. Technological, organizational, and environmental antecedents to web services adoption. Commun IIMA 2006;6(1):147-160 [FREE Full text]
68. Xu S, Zhu K, Gibbs J. Global technology, local adoption: a cross-country investigation of internet adoption by companies in the United States and China. Electron Mark 2004;14(1):13-24. [doi: 10.1080/1019678042000175261]
69. Bouchard L. Decision Criteria in the Adoption of EDI. In: Proceedings of the 14th International Conference on Information Systems. 1993 Presented at: ICIS'93; December 5-8, 1993; Orlando, Florida, USA p. 365-376 URL: https://pdfs.semanticscholar.org/1900/4658d2e7b34f7ae521a37888da458c6eb2ab.pdf
70. Hart PJ, Saunders CS. Emerging electronic partnerships: antecedents and dimensions of EDI use from the supplier's perspective. J Manag Inf Syst 1998;14(4):87-111 [FREE Full text] [doi: 10.1080/07421222.1998.11518187]
71. Premkumar G, Ramamurthy K. The role of interorganizational and organizational factors on the decision mode for adoption of interorganizational systems. Decis Sci 1995 May;26(3):303-336. [doi: 10.1111/j.1540-5915.1995.tb01431.x]
72. Morison E. Gunfire at sea: a case study of innovation. In: Tushman M, Anderson P, editors. Managing Strategic Innovation and Change. New York: Oxford University Press; 1997:129-140.
73. Thatcher SM, Foster W, Zhu L. B2B e-commerce adoption decisions in Taiwan: The interaction of cultural and other institutional factors. Electron Commer Res Appl 2006 Jun;5(2):92-104. [doi: 10.1016/j.elerap.2005.10.005]
74. Moon MJ, Bretschneider S. Can state government actions affect innovation and its diffusion?: An extended communication model and empirical test. Technol Forecast Soc Change 1997 Jan;54(1):57-77. [doi: 10.1016/s0040-1625(96)00121-7]
75. Kraemer KL, Gurbaxani V, King JL. Economic development, government policy, and the diffusion of computing in Asia-Pacific countries. Public Adm Rev 1992 Mar;52(2):146-156. [doi: 10.2307/976468]
76. Rosenberg N. Inside The Black Box. Cambridge: Cambridge University Press; 1982.
77. Arthur WB. Competing technologies: an overview. In: Dosi G, Freeman C, editors. Technical Change and Economic Theory. London: Pinter Publishers; 1988:590-607.
78. Farrell J, Saloner G. Competition, compatibility, and standards: the economics of horses, penguins and lemmings. In: Gabel HL, editor. Product Standardization and Competitive Strategy. Amsterdam: Elsevier Science; 1986:1-21.
79. Premkumar G, Ramamurthy K, Nilakanta S. Implementation of electronic data interchange: an innovation diffusion perspective. J Manag Inf Syst 1994;11(2):157-186 [FREE Full text] [doi: 10.1080/07421222.1994.11518044]
80. Rogers EM. Diffusion of Innovations. Fifth Edition. New York: Free Press; 2003.
81. Frambach RT. An integrated model of organizational adoption and diffusion of innovations. Eur J Mark 1993 Jun;27(5):22-41. [doi: 10.1108/03090569310039705]
82. Gharavi H, Love PE, Cheng EW. Information and communication technology in the stockbroking industry: an evolutionary approach to the diffusion of innovation. Indus Manag Data Syst 2004 Dec;104(9):756-765. [doi: 10.1108/02635570410567748]
83. King WR, Malhotra Y. Developing a framework for analyzing IS sourcing. Inf Manag 2000 Sep;37(6):323-334. [doi: 10.1016/s0378-7206(00)00046-x]
84. Yang C, Huang J. A decision model for IS outsourcing. Int J Inf Manag 2000 Jun;20(3):225-239. [doi: 10.1016/s0268-4012(00)00007-4]
85. Gottardi G, Bolisani E, Di Biagi M. Electronic commerce and open communities: an assessment of internet EDI. Int J Serv Technol Manag 2004;5(2):151-169. [doi: 10.1504/ijstm.2004.004056]
86. Ouchi W, Williamson OE. Markets and hierarchies: analysis and antitrust implications. Adm Sci Q 1977 Sep;22(3):540. [doi: 10.2307/2392191]
87. Zand DE. Trust and managerial problem solving. Adm Sci Q 1972 Jun;17(2):229-239. [doi: 10.2307/2393957]
88. Shang RA, Chen CC, Liu YC. Internet EDI Adoption Factors: Power, Trust and Vision. In: Proceedings of the 7th international conference on Electronic commerce. 2005 Presented at: ICEC'05; August 15-17, 2005; Xi'an, China p. 101-108. [doi: 10.1145/1089551.1089573]
89. Levine S, White PE. Exchange as a conceptual framework for the study of interorganizational relationships. Adm Sci Q 1961 Mar;5(4):583-601. [doi: 10.2307/2390622]

90. Mohr J, Spekman R. Characteristics of partnership success: Partnership attributes, communication behavior, and conflict resolution techniques. Strat Manag J 1994 Feb;15(2):135-152. [doi: 10.1002/smj.4250150205]
91. Frazier GL, Spekman RE, O'Neal CR. Just-in-time exchange relationships in industrial markets. J Mark 1988;52(4):52-67. [doi: 10.1177/002224298805200406]
92. Ratnasingam P. The influence of power on trading partner trust in electronic commerce. Internet Res 2000 Mar;10(1):56-63. [doi: 10.1108/eum0000000005316]
93. Ke W, Liu H, Wei KK, Gu J, Chen H. How do mediated and non-mediated power affect electronic supply chain management system adoption? The mediating effects of trust and institutional pressures. Decis Sup Syst 2009 Mar;46(4):839-851. [doi: 10.1016/j.dss.2008.11.008]

Abbreviations

HIT: health information technology
IT: information technology
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RQ: research question
TOE: technology-organization-environment

Edited by C Lovis; submitted 07.12.19; peer-reviewed by U Taneja, Y Tani, A Rezaei Aghdam; comments to author 01.02.20; revised version received 16.02.20; accepted 20.02.20; published 15.05.20.

Please cite as:
Han L, Liu J, Evans R, Song Y, Ma J
Factors Influencing the Adoption of Health Information Standards in Health Care Organizations: A Systematic Review Based on Best Fit Framework Synthesis
JMIR Med Inform 2020;8(5):e17334
URL: https://medinform.jmir.org/2020/5/e17334
doi:10.2196/17334
PMID:32347800

Review

Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping

Elsie Horne[1], BSc, MSc; Holly Tibble[1], BSc, MPhil(Cantab); Aziz Sheikh[1], BSc, MSc, MBBS, MD; Athanasios Tsanas[1], BSc, BEng, MSc, DPhil(Oxon)

Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, United Kingdom

Corresponding Author:
Elsie Horne, BSc, MSc
Usher Institute, Edinburgh Medical School
University of Edinburgh
Nine Edinburgh Bio Quarter
9 Little France Road
Edinburgh, EH16 4UX
United Kingdom
Phone: 44 1316517887
Fax: 44 1316517887
Email: Elsie.Horne@ed.ac.uk

Abstract

Background: In the current era of personalized medicine, there is increasing interest in understanding the heterogeneity in disease populations. Cluster analysis is a method commonly used to identify subtypes in heterogeneous disease populations. The clinical data used in such applications are typically multimodal, which can make the application of traditional cluster analysis methods challenging.

Objective: This study aimed to review the research literature on the application of clustering multimodal clinical data to identify asthma subtypes. We assessed common problems and shortcomings in the application of cluster analysis methods in determining asthma subtypes, such that they can be brought to the attention of the research community and avoided in future studies.

Methods: We searched PubMed and Scopus bibliographic databases with terms related to cluster analysis and asthma to identify studies that applied dissimilarity-based cluster analysis methods. We recorded the analytic methods used in each study at each step of the cluster analysis process.

Results: Our literature search identified 63 studies that applied cluster analysis to multimodal clinical data to identify asthma subtypes. The features fed into the cluster algorithms were of a mixed type in 47 (75%) studies and continuous in 12 (19%), and the feature type was unclear in the remaining 4 (6%) studies. A total of 23 (37%) studies used hierarchical clustering with Ward linkage, and 22 (35%) studies used k-means clustering. Of these 45 studies, 39 had mixed-type features, but only 5 specified dissimilarity measures that could handle mixed-type features. A further 9 (14%) studies used a preclustering step to create small clusters to feed on a hierarchical method. The original sample sizes in these 9 studies ranged from 84 to 349. The remaining studies used hierarchical clustering with other linkages (n=3), medoid-based methods (n=3), spectral clustering (n=1), and multiple kernel k-means clustering (n=1), and in 1 study, the methods were unclear. Of 63 studies, 54 (86%) explained the methods used to determine the number of clusters, 24 (38%) studies tested the quality of their cluster solution, and 11 (17%) studies tested the stability of their solution. Reporting of the cluster analysis was generally poor in terms of the methods employed and their justification.

Conclusions: This review highlights common issues in the application of cluster analysis to multimodal clinical data to identify asthma subtypes. Some of these issues were related to the multimodal nature of the data, but many were more general issues in the application of cluster analysis. Although cluster analysis may be a useful tool for investigating disease subtypes, we recommend that future studies carefully consider the implications of clustering multimodal data, the cluster analysis process itself, and the reporting of methods to facilitate replication and interpretation of findings.

(JMIR Med Inform 2020;8(5):e16452) doi:10.2196/16452

KEYWORDS

asthma; cluster analysis; data mining; machine learning; unsupervised machine learning

Introduction

Background

There is mounting evidence to suggest that some disease labels are in fact *umbrella terms*, which encompass distinct disease subtypes with different underlying mechanisms and clinical symptom manifestations [1-3]. This has encouraged the investigation into heterogeneity within disease populations, which has received considerable interest across diverse domains of medicine [4-6]. There are numerous motivations for better understanding heterogeneity within disease populations, from the development of targeted therapeutics [6] to the delivery of more personalized care in clinical practice [7].

It is now understood that asthma is one such umbrella term used to encompass multiple diverse underlying disease symptoms and pathophysiology [7]. Asthma is a common chronic condition characterized by reversible airway obstruction. The Global Burden of Disease Study 2017 estimated the global prevalence of asthma (both symptomatic and asymptomatic) to be 273 million [8]. This study estimated that in 2017, there were 43 million new cases of asthma and 495,000 deaths attributed to asthma [9]. Attempts to categorize asthma into distinct disease subtypes date back to the 1940s [10] and are ongoing. However, the methods for discovering these underlying categories have shifted from observing clinical patterns to using data-driven approaches such as *cluster analysis* [11].

Cluster analysis is a statistical technique used to identify subgroups in data based on multiple variables (for convenience, herein, we have used the term *features*). It is an *unsupervised* statistical learning method, and the correct number of underlying clusters is typically unknown *a priori* [12]. The technique has found increasing use in recent years because of the practical unmet clinical need to identify subtypes of disease and stratify patients to improve health care delivery. This has been made feasible by the increasing availability of clinical datasets and the development of statistical software packages facilitating the application of algorithmic methods.

Clinical datasets are often *multimodal*; for the purposes of this paper, we defined a multimodal dataset as a dataset that includes features from different sources, measured on different scales. For completeness and to avoid ambiguity, we clarified that the term multimodal has a different meaning in statistical literature (ie, features with multiple modes in terms of its distribution); the use of the term in this study is aligned with clinical literature (having features from different sources). Popular methods of cluster analysis such as k-means and hierarchical clustering with the Ward method have been developed for continuous features measured on a common scale. In practice, however, many of these techniques are frequently applied to multimodal clinical datasets comprising different feature types measured on different scales, conditions that violate some of the underlying principles and assumptions made by algorithmic methods [13]. Although steps can be taken to prepare multimodal clinical data for cluster analysis [13], the results of a previous review suggest that these steps are rarely taken in practice [11]. This previous review focused on the clinical findings of the studies and touched only briefly on the challenges of clustering multimodal data specifically.

Objectives

This review aimed to comprehensively explore whether studies applying cluster analysis to multimodal clinical data to subtype asthma are using appropriate clustering methodologies. The contribution of this study is to make recommendations for the robust application of cluster analysis to multimodal clinical data. We believed this would be of interest to the ever-growing number of asthma researchers engaging or planning to engage in disease subtyping, as well as to the wider community of researchers applying cluster techniques for the purpose of disease subtyping.

Methods

Eligibility Criteria and Search Strategy

This review is reported following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Multimedia Appendix 1 shows the completed PRISMA checklist.

We sought to identify studies that applied cluster analysis to multimodal clinical data with the aim of identifying subtypes of asthma. One researcher (EH) searched PubMed and Scopus databases (search queries are provided in Textbox 1) to retrieve studies focusing on patients diagnosed with asthma, which included the term *cluster analysis* or *clustering*. Our search was restricted to studies published between January 1, 2008, and May 23, 2019, as Haldar et al's study [14] is widely acknowledged to be the first to apply cluster analysis to identify subtypes of asthma. Our search excluded comment articles, editorials, letters, reviews, and meta-analyses. We excluded articles that were not written in English.

We excluded nonrelevant studies by first screening the abstracts, then referring to the full text where necessary. We excluded studies in which (1) none of the aims or objectives were to identify subtypes of asthma (studies looking exclusively at, eg, childhood wheeze were excluded); (2) the data were not multimodal (ie, were measured from a common source and on a common scale); and (3) none of the features were considered clinical (eg, studies concerned only with -omics data). Finally, we excluded studies that used latent class analysis or mixture models to group their data to narrow the scope of this review to methods that cluster samples based on pairwise dissimilarities. The use of latent class analysis to distinguish asthma phenotypes has been reviewed previously by Howard et al [15].

Textbox 1. Search query to identify studies to include in this review.

- The following query was inserted in PubMed on May 23, 2019:

  English[Language] AND ("2008/01/01"[Date - Publication] : "2019/05/23"[Date - Publication]) AND ("cluster analysis"[Text Word] OR "clustering"[Text Word]) AND "asthma*"[Text Word] NOT (comment[Publication Type] OR editorial[Publication Type] OR letter[Publication Type] OR review[Publication Type] OR meta-analysis[Publication Type])*

- The following query was inserted in Scopus on May 23, 2019:

  PUBYEAR > 2007 AND (TITLE-ABS-KEY ("cluster analysis") OR TITLE-ABS-KEY("clustering")) AND TITLE-ABS-KEY ("asthma*") AND SRCTYPE ("j") AND DOCTYPE ("ar") AND LANGUAGE ("English")*

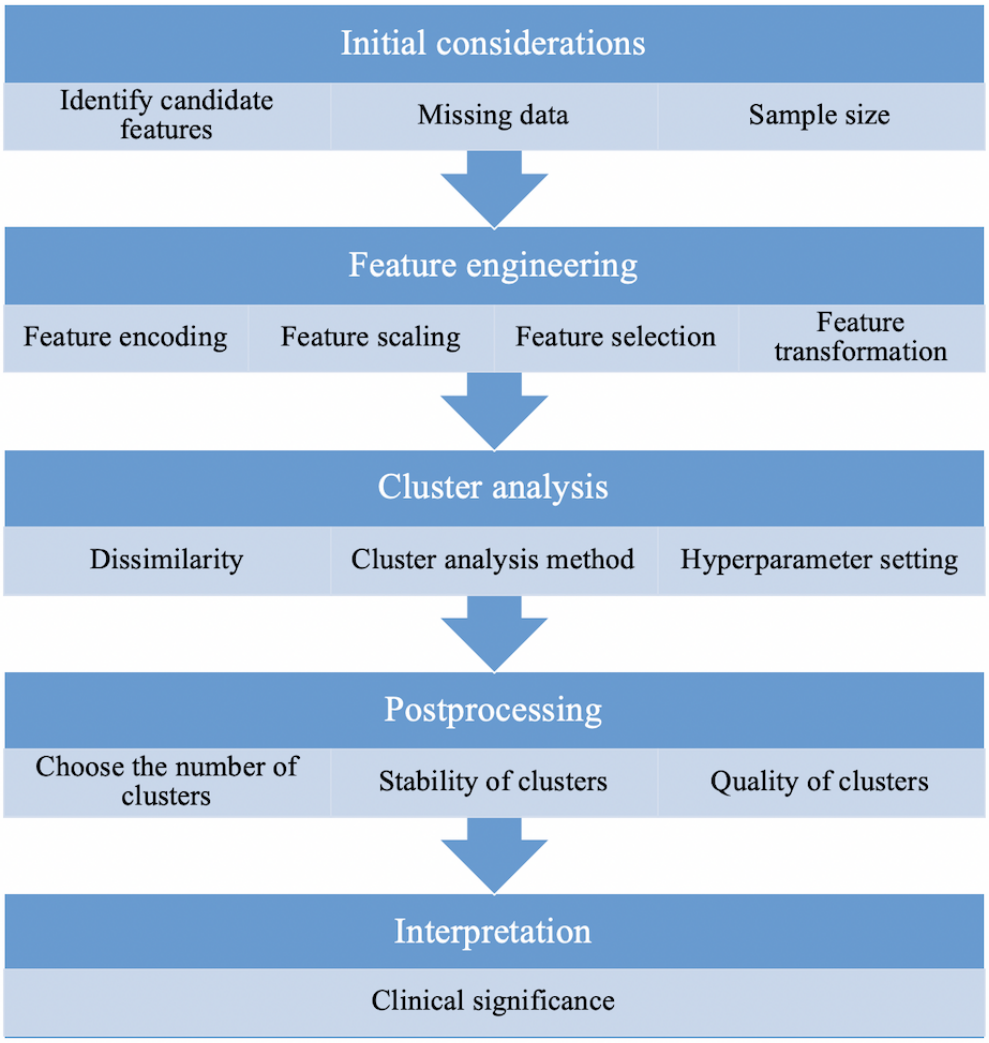
Data Extraction

In total, 2 researchers (EH and HT) independently extracted information from the full text and supplementary material of each study. Information was extracted following the steps outlined in the following *Cluster Analysis Steps* section. The data dictionary, which provides details of all items extracted, is presented in Multimedia Appendix 2.

Cluster Analysis Steps

To provide context for this review, we outlined the key steps in the application of cluster analysis to multimodal clinical data. Figure 1 summarizes the steps in the order in which they generally occur, but as with most analytic processes, this depends on the context, and the process may be somewhat iterative.

Figure 1. Schematic of the typical cluster analysis steps.



Initial Considerations

Identify Candidate Features

The first step is to identify the set of features of interest, which we referred to as *candidate features*. These may be identified based on previous studies or clinical input using domain expertise. In some cases, all the candidate features may be used in the cluster analysis (we referred to the features used in cluster analysis as *cluster features*). In other cases, formal feature selection processes may be applied to the candidate features to identify the cluster features, as covered in the *Feature Selection* section.

Missing Data

Most common cluster analysis methods use *complete case analysis* (ie, the cluster features have no missing entries, which, in practice, might be achieved by removing samples for which any cluster feature entry is missing). However, it may be more data efficient to develop a strategy to work around missing entries instead of discarding samples. Missing values may be handled through the calculation of dissimilarities, as described by Hastie et al [16]. Alternatively, missing data could be imputed, or for categorical features, a missing category could be introduced.

Sample Size

Despite the widespread use of cluster analysis, at present, there is no consensus regarding the minimum sample size required to ensure stable and meaningful clustering. Dolnicar et al [17] suggested that 70 samples per cluster feature is adequate, based on the findings of their simulation study. Small sample sizes may obscure the true clustering by causing the user to pick the wrong number of clusters (see the *Choosing the Number of Clusters* section) or by producing solutions that are neither reproducible nor stable (see the *Stability* and *Quality* subsections).

Feature Engineering

Feature Types

The features that we may want to use in a clustering algorithm often come from multimodal clinical data. Hence, they may be of different types (eg, continuous, nominal, ordinal, binary, etc) and are likely to be measured on different scales (eg, kilogram for mass, years for age). Most dissimilarity measures and clustering algorithms assume that the features are of the same type and are measured on a common scale. These requirements can be addressed using *feature encoding* and *feature scaling*.

Feature Encoding

When dealing with categorical features, it is vital to consider how these are encoded (nominal, ordinal, or binary), as this determines how they are treated in the calculation of dissimilarities and in the clustering algorithm. A common approach is to encode ordinal features as integers and to encode nominal features as dummy binary features [18].

Feature Scaling

Feature scaling may be used to address 3 issues related to continuous features. The first is that continuous features may be measured in different units and should therefore be rescaled to bring them onto a common scale before calculating dissimilarities. The second is that continuous features measured in the same units may have different variances. In some cases, the differences in variance may be useful for clustering, but in others, these may obscure the true underlying cluster structure in the data. In the latter case, the continuous features should be rescaled. Common approaches to these 2 issues are to standardize features to have 0 mean and unit variance (referred to as *z-scores*) or to use range normalization techniques, for example, to scale each feature so that it is in the interval of 0 to 1.

The third issue is that the features may not follow the desired probability distribution properties for further analysis (eg, having Gaussian-distributed features). This issue needs to be considered when statistical methods make distributional assumptions. Although few dissimilarity-based clustering methods make distributional assumptions, several methods involve the calculation of cluster means (eg, k-means, hierarchical clustering with the Ward linkage). The mean is a poor choice of summary statistic for a feature that is skewed (or a feature with multiple modes), so a power transformation may be advantageous as a preprocessing step when using such clustering methods.

When dealing with mixed-type data, it may be necessary to scale the categorical features to avoid assigning categorical features greater weight over continuous features or vice versa. This issue is discussed in detail in the context of dissimilarity measures by Hennig and Liao [13].

Dimensionality Reduction

There are generally 2 motivations for reducing the dimensionality of a dataset before applying cluster analysis. First, as previously mentioned in the *Sample Size* subsection, datasets with a high feature to sample ratio may not produce stable cluster results. Second, the cluster structure may only be apparent using a subset of the information available in the data. Using all available information may introduce noise, which could obscure the true underlying cluster structure [19]. There are 2 approaches to dimensionality reduction: *feature selection* and *feature transformation*.

Feature Selection

Feature selection involves selecting a subset of the available features for use in cluster analysis. Herein, we have referred to the features selected for the cluster analysis as *cluster features*.

Feature Transformation

Feature transformation involves combining original features to create new features. Generally, a subset of these new features is selected for inclusion in the analysis. It is beyond the scope of this review to provide in-depth details on the methods of feature transformation (also known as *feature extraction*); we referred to van der Maaten et al's [20] work for a comprehensive review. Here, we briefly outlined *principal component analysis* (PCA), which is the most commonly used method for linear data projection. PCA may be applied to p continuous, correlated features to extract $m<p$ continuous, and uncorrelated features (known as *principal components*), each being a linear function of the original cluster features [21]. Related methods include factor analysis for continuous data, *multiple correspondence*

analysis (MCA) for categorical data [22], and multiple factor analysis for mixed-type data [23].

Cluster Analysis

Dissimilarity Measures

Model-free clustering methods rely on a *dissimilarity measure* to quantify how dissimilar 2 samples are from one another. Dissimilarity may also be referred to as a *distance measure* if it satisfies the triangle inequality. The most widely used dissimilarity measure is the squared Euclidean distance (henceforth referred to as *Euclidean distance*), which is intended for use with continuous features. A dissimilarity measure that can handle both categorical and continuous features is the Gower distance [24].

Cluster Analysis Methods

There are many different methods of cluster analysis (eg, k-means, hierarchical clustering with the Ward linkage, spectral clustering), and each method may be implemented using different algorithms. A comprehensive overview of the wide range of clustering methods can be found elsewhere [25].

Postprocessing

Choosing the Number of Clusters

A key challenge in cluster analysis is choosing the number of clusters to present in the final solution, which is typically unknown *a priori*. Often, researchers use their preferred clustering methods, running them for 2 to *k* clusters (where *k* is an integer number indicating the number of clusters) and then have a strategy to determine *k*.

Providing a detailed commentary on these strategies is beyond the scope of this review. An overview of strategies for choosing *k* is provided by Everitt et al [23]. Graphical techniques include dendrograms (when using hierarchical clustering methods) and silhouette plots [26]. An alternative approach is to choose the number of clusters that gives the most stable solution [27]. In practice, a key determinant in choosing the number of clusters is often the clinical interpretation of the solutions.

We highlighted the possibility that there might not be meaningful clustering of the data to form groups, and thus, the entire dataset is treated as 1 cluster. This may reflect the lack of statistical power (sufficiently large sample size) to determine clusters or that the investigated problem using that dataset is not amenable to clustering using the available sample size and features. Some statistics used for choosing k, such as the Gap statistic [28], can be calculated for k=1. However, statistics that require the calculation of between cluster differences or distances, such as the silhouette statistic, are not defined for k=1 [26].

Stability

Assessing the quality of a clustering solution produced using any cluster algorithm is challenging. Unlike supervised learning setups, there is no *ground truth* against which one can formally test their findings. However, there are several ways in which one can assess the integrity of their findings.

Most importantly, it is crucial to assess the *stability* of the resulting clusters. A definition of *cluster stability*, given by von Luxburg [27], is whether clustering different datasets sampled from the same underlying joint distribution will result in producing the same clusters. There are several ways in which this may be assessed in practice (eg, by comparing the cluster results of a dataset that has been randomly split into 2 or more subsets, and each subset is independently fed into the cluster algorithm).

Quality

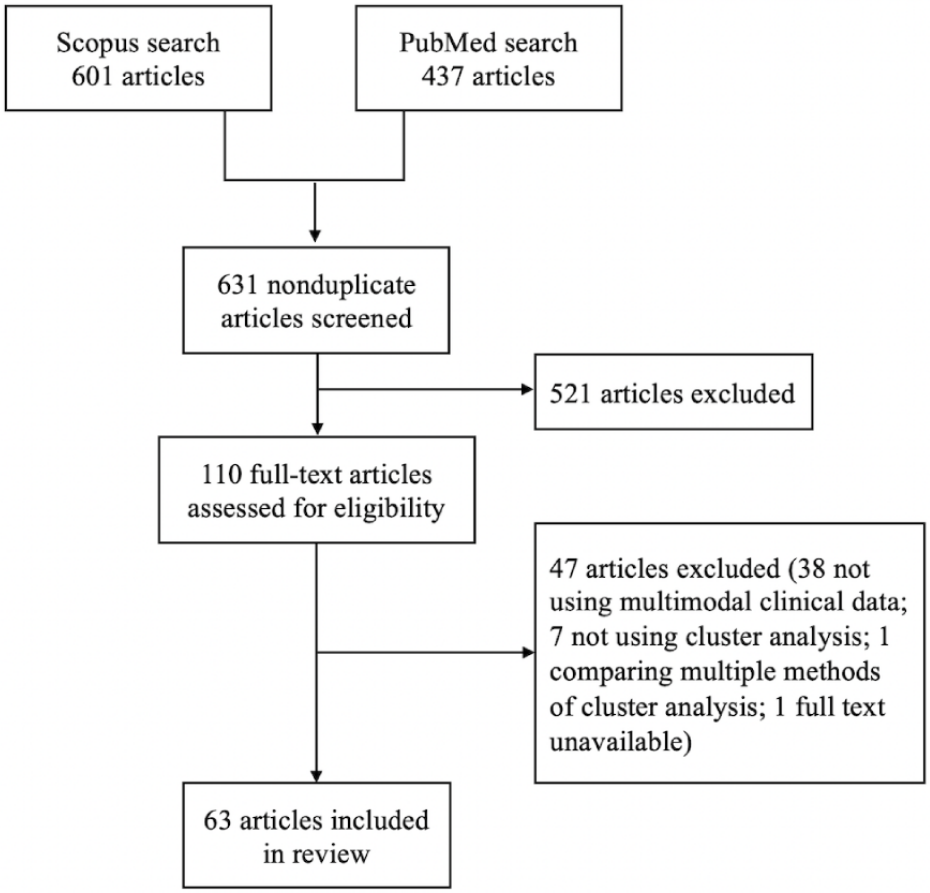
Beyond stability, there are numerous steps one may take to ensure the integrity of their cluster analysis findings, for example, repeating the analysis in a different cohort or at a different time point, or altering the encoding of a feature. These steps are often referred to as reproducibility testing. However, we avoided this term because it implies that we seek the exact same results, which we do not feel is reasonable in all scenarios. To extract this information from the studies in this review, 2 reviewers independently extracted details of postprocessing methods, which we felt assessed the quality of the cluster results, but did not come under stability. In our schematic and results, we referred to these methods as testing the quality of the cluster results.

Results

Literature Search Outcomes

We identified 63 studies that used cluster analysis to identify subtypes of asthma using multimodal clinical data (Figure 2). One of the excluded articles satisfied our inclusion criteria but investigated 85 combinations of cluster analysis steps in a hierarchical cluster analysis of 383 children with asthma [29]. We excluded this study from our review as including all 85 combinations of methods was deemed infeasible. For the 2 studies in which cluster analysis was carried out in multiple populations [14,28], we included only the analysis of the larger population. The characteristics of each study are presented in Multimedia Appendix 3.

Figure 2. Flow of studies into review.



Initial Considerations

Identifying Candidate Features

A total of 42 (67%) studies identified candidate features based on previous studies or clinical input (relevance to asthma subtypes, avoiding clinical redundancy, and easily measured in clinical practice). The numbers used in each method are summarized in Table 1.

Missing Data

A total of 42 (67%) studies detailed their methods for dealing with missing data; the methods used are shown in Table 1. The most common method was to carry out a complete case analysis by excluding all patients with any missing cluster feature entries (35% of studies).

Table 1. Initial considerations across the asthma studies we have included in this review (N=63).

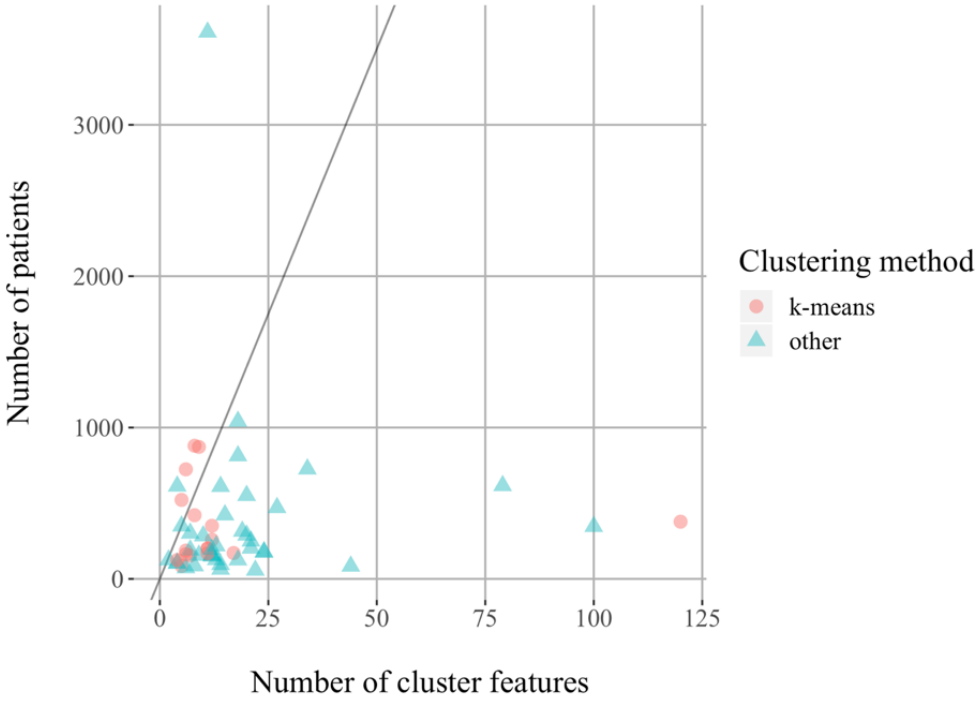
Method	Values, n (%)[a]
Identifying candidate features	
Clinical intuition and understanding	33 (52)
Avoid clinical redundancy	15 (24)
Previous studies	15 (24)
Easily measured in clinical practice	8 (13)
Missing data	
Complete case analysis	22 (35)
Features with >x%[b] missing values removed	14 (22)
Imputed	11 (17)
Patients with >x%[b] missing values removed	5 (8)
No missing data present	2 (3)
Clustering methods handle missing data	1 (2)

[a]One study may use multiple methods; some studies may use no methods.
[b]x>0.

Sample Size

The sample sizes for cluster analysis ranged from 40 to 3612, with a median of 195 patients. Figure 3 presents a scatter plot of the number of patients in the cluster analysis versus the final number of cluster features. The straight line corresponds to the number of samples per feature as recommended by Dolnicar et al [17]. As this estimate was derived from simulation studies using k-means as the clustering method, different markers are used for the studies which used clustering techniques other than k-means. Note that the studies that did not specify the final number of cluster features were omitted from the plot. Six studies (10%) had at least 70 times as many patients as cluster features, as recommended by Dolnicar et al [17].

Figure 3. Number of patients versus final number of cluster features. The line corresponds to the number of patients that is equal to 70 times the number of features.



Feature Engineering

Feature Scaling and Encoding

Judging whether feature scaling and encoding were appropriate depends on the methods of cluster analysis used and vice versa. Therefore, we reported the methods of feature scaling and encoding alongside the methods of cluster analysis in Tables 2-4 and Multimedia Appendix 4.

Table 2. Breakdown of methods used by studies applying hierarchical clustering with Ward's linkage (N=23).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
Not detailed	N/A[a]	1 (4)
Mixed		
Euclidean assumed		
Scaled but method unspecified	Yes	1 (4)
	No	1 (4)
Scaled to lie in the interval of 0 to 1	Yes	1 (4)
z-scores	Yes	1 (4)
	No	1 (4)
Not detailed	Yes	3 (13)
	No	6 (26)
Euclidean stated		
z-scores	Yes	2 (9)
	No	1 (4)
Gower[b]		
Gower standardisation	No	3 (13)
Scaled but method unspecified	No	1 (4)
treeClust		
Not detailed	No	1 (4)

[a]N/A: not applicable (irrelevant for continuous features).

[b]Computing the Gower coefficient normalizes the distance between feature samples by dividing by the feature range. Therefore, it is not necessary to normalize continuous features prior to computing the Gower coefficient.

Table 3. Breakdown of methods used by studies applying k-means (N=22).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
z-scores for one feature	N/A[a]	1 (5)
No details	N/A	3 (14)
Euclidean stated		
No details	N/A	1 (5)
Mixed		
Euclidean assumed		
Scaled but method unspecified	No	1 (5)
z-scores	Yes	6 (27)
z-scores for one feature	No	1 (5)
No details	Yes	1 (5)
	No	2 (9)
Euclidean stated		
z-scores	Yes	1 (5)
No details	No	1 (5)
Unclear		
Euclidean assumed		
No details	No	3 (14)
Euclidean stated		
z-scores	No	1 (5)

[a]N/A: not applicable (irrelevant for continuous features).

Table 4. Breakdown of methods used by studies applying SPSS TwoStep (N=7).

Data type, dissimilarity, and scaling of continuous features	Categorical features encoded as binary?	Value, n (%)
Continuous		
Euclidean assumed		
No details	N/A[a]	1 (14)
Mixed		
Log-likelihood assumed		
Scaled to lie in the interval 0 to 1	Yes	1 (14)
z-scores	No	1 (14)
No details	Yes	2 (29)
Log-likelihood stated		
Scaled but method unspecified	No	1 (14)
No details	No	1 (14)

[a]N/A: not applicable (irrelevant for continuous features).

Univariate Feature Transformation

A total 23 (37%) studies applied univariate feature transformation to bring features closer to a normal distribution. The most common univariate feature transformation was logarithmic transformation, applied to nonnormally distributed features in 33% of studies. Lefaudeux et al [30] applied the Box-Cox transformation to all features, whereas Khusial et al [31] stated that data were transformed if necessary but gave no further details.

Feature Selection

A total of 22 (35%) studies detailed methods of feature selection to identify their cluster features. The number of features selected in the 63 studies included in this review ranged from 2 to 120, with a median of 12 features. In addition, 47 (75%) studies had mixed-type features, and 12 (19%) had continuous features, and in 4 (6%) studies, the type of features was unclear. Methods for feature selection are listed in Table 5.

A total of 13 (20%) studies used PCA or factor analysis for feature selection. These are not typically methods that should be used for feature selection; we defer further elaboration on the topic for the Discussion. All but one of these studies computed the components (or factors) that represent an underlying latent feature structure, then selected 1 (or in some cases multiple [32,33]) original feature corresponding to each component (or factor) of the latent feature structure. Just et al [34] stated that they used PCA to select features according to statistical significance. As PCA does not involve the computation of statistical significance (*P* values), more detail would be required here to fully understand the methods used for feature selection in this paper. Pérez-Losada et al [35] stated PCA based on Euclidean distances was carried out. It is unclear whether this was an error in reporting or whether PCA was applied to the matrix of Euclidean distances between features instead of the covariance matrix. To implement the latter approach, the Euclidean distances would have to be converted to similarities. Moreover, the authors stated that PCA was used *to identify key clinical components relevant to asthma diagnosis and assessment*. Overall, it is not clear how the authors processed the data using PCA, and there was no justification for using Euclidean distances in that computation. Although the application of PCA leads to the computation of features (principal components) that maximally explain the (remaining) variance in the data, there is no guarantee that the resulting principal components will be highly predictive of an outcome (in this case, asthma diagnosis and assessment).

Table 5. Feature engineering methods used in the asthma studies included in this review.

Method	Values, n (%)[a]
Univariate feature transformation	
Logarithmic transformation	21 (33)
Box-Cox transformation	1 (2)
Method not explained	1 (2)
Feature selection	
Factor analysis[b]	8 (13)
Principal component analysis[b]	5 (8)
Avoid collinearity	3 (5)
Avoid multicollinearity	3 (5)
Supervised learning methods	2 (3)
Multiple correspondence analysis	1 (2)
Feature transformation	
Principal component analysis	4 (6)
Factor analysis	1 (2)
Multiple correspondence analysis	1 (2)

[a]As a percentage of all 63 studies.

[b]These are not typically methods of feature selection but have been used in these studies.

Three (5%) studies considered collinearity via pairwise correlations, although the exact criteria for selection features based on this were unclear [36-38]. In addition, 3 (5%) studies avoided multicollinearity, but none detailed their methods for doing so [39-41].

Furthermore, 2 (3%) studies selected features using statistical hypothesis tests with respect to the outcome of interest. Sakagami et al [42] used mean annual decline in forced expiratory volume in 1 second as the outcome feature in a multiple regression analysis using stepwise feature selection. All features with coefficients statistically significantly different to 0 in the multiple regression model were included as cluster features. Seino et al [43] grouped participants according to whether or not they had symptoms of depression. Features were selected for cluster analysis if the difference between the 2 groups (tested using a Wilcoxon rank-sum or chi-square test for continuous and categorical features, respectively) was statistically significant.

Feature Transformation

A total of 6 (10%) studies performed feature transformation before cluster analysis; the methods are summarized in Table 5. Of the 4 studies that used PCA for feature transformation, 3 used continuous input features [30,44,45], whereas the fourth used mixed-type input features [46]. None of the studies stated whether the covariance or correlation matrix was used as input

for PCA. Only Newby et al [45] specified the number of transformed features retained, and the proportion of original variance accounted for.

Khusial et al [31] performed factor analysis on a subset of the selected features; it is unclear whether categorical features are included in this subset. Although the resulting factors were scaled to z-scores, the authors did not provide further information regarding whether the features were scaled before factor analysis. Four factors were retained, but neither the proportion of variance explained by these factors nor a table of the factor loadings is given.

Sendín-Hernández et al [47] performed MCA to transform 5 continuous and 14 categorical features. They gave the proportion of variance explained by the transformed features but gave neither the number of transformed features retained nor a table of the feature loadings.

Cluster Analysis

Hierarchical Clustering

A total of 23 (37%) studies applied hierarchical clustering with the Ward method [48] as the principal clustering technique. A breakdown of the methods used by these studies is given in Table 2. One study applied these methods to continuous data, and the remaining 22 studies used mixed-type data. Three studies stated that the Euclidean distance was used, 4 used Gower coefficient (issues with the Gower coefficient combined with the Ward method are addressed in the *Discussion* section), and 1 used tree-based dissimilarity measure [49]. For the remaining 15 studies, we assumed that the Euclidean distance was used. Of the 23 studies, 11 did not detail whether the features were rescaled. Of the 17 studies using the Euclidean distance with mixed-type features, 8 encoded categorical features as binary features.

A total of 3 (5%) further studies (in addition to the 23 studies introduced at the start of the paragraph) applied hierarchical clustering to continuous data. Amore et al [39] used the average linkage and the Euclidean distance, whereas 2 studies used hierarchical clustering but did not specify the linkage or dissimilarity measure used [44,50].

k-Means

A total of 22 (35%) studies used k-means clustering as the principal clustering technique. A breakdown of the methods used by these 3 studies is given in Multimedia Appendix 4. A breakdown of the methods used by these studies is given in Table 3. Five studies applied k-means to continuous data, and 13 studies applied it to mixed-type data. In 3 studies, the cluster features were not explicitly stated, and the data types therefore were unclear. Of the 22 studies, 4 explicitly stated that the Euclidean distance was used. As no other dissimilarity metrics were mentioned, we assumed that the Euclidean distance was used in the remaining 18 studies because it is often the default option for most algorithmic packages. Of the 22, 11 studies did not detail whether continuous features were scaled before cluster analysis. Of the 13 studies with mixed-type data, 8 encoded categorical features as binary features.

Preclustering Methods

When dealing with very large sample sizes, it can be advantageous to introduce a precluster step. The aim is to group samples and to use these groups or *preclusters* as input to a follow-on clustering algorithm (ie, using 2 steps with cascaded cluster algorithms). This step is used to reduce the computation time required to compute the cluster results.

A total of 7 (11%) studies used the SPSS TwoStep clustering method [51,52]. A breakdown of the preprocessing methods and distance measures used by these studies is given in Table 4. In the first (precluster) step, a cluster feature tree is identified. In the second step, the preclusters are merged stepwise until all clusters are in 1 cluster using the Euclidean or log-likelihood distance for continuous or mixed-type features, respectively. An advantage of the log-likelihood distance measure is that it is designed to handle mixed-type features. However, in doing so, it assumes that continuous (categorical) features follow a normal (multinomial) distribution within clusters.

None of the studies in this review adequately considered the distributional assumptions made by the SPSS TwoStep method. Ruggieri et al [53] acknowledged that the method assumes continuous features are normally distributed, but they did not explicitly report whether these assumptions were satisfied. Although Newby et al [45] acknowledged that the method assumes cluster features are statistically independent within clusters, they only go as far as to ensure that their cluster features are uncorrelated (by applying PCA), which does not necessarily imply independence. The remaining 5 studies that used the SPSS TwoStep method did not reference distributional assumptions.

Two (3%) further studies preclustered samples (Just et al [34] specified k-means, and Ye et al [54] did not specify the precluster method) and then applied hierarchical clustering with the Ward linkage method on the preclusters. A breakdown of the methods used by these 2 studies is given in Multimedia Appendix 4.

k-Medoid Methods

Three studies used k-medoid methods. A breakdown of the methods used by these 3 studies is given in Multimedia Appendix 4. Two used k-medoids implemented by the Partition Around Medoids algorithm [55]. Lefaudeux et al [30] used the Euclidean distance with center-scaled continuous data, and Sekiya et al [56] used the Gower metric with mixed-type data. Loza et al [57] applied fuzzy partition-around-medoid clustering with the Euclidean distance to continuous data scaled with average absolute deviation.

Kernel k-Means and Spectral Clustering

Kernel k-means and spectral clustering are different but related methods, which may be used to identify clusters that are not linearly separable in the input feature space [58]. As these methods were used by only 1 study each (Wu et al used multiple kernel k-means [59], and Howrylak et al used spectral clustering [37]), we do not explore them in detail in this review. However, details of the feature scaling, encoding, and distance measures used by these 2 studies is given in Multimedia Appendix 4.

Unclear Methods

Wang et al [41] described a 2-step clustering method in which the first step was to carry out hierarchical clustering using the Ward method, but with the log-likelihood distance in place of the Euclidean distance. This first step was used to determine the number of clusters, which was then used in the k-means method in the second step. However, the authors cite the SPSS TwoStep method [52], which is different from that described previously. It was therefore ambiguous which clustering method was applied in this study.

Postprocessing

Choosing the Number of Clusters

A total of 54 (86%) studies explained in detail the methods used to select the number of clusters. Of these, 20 (32%) studies used more than one method for choosing the number of clusters. The maximum number of methods used was 6.

A total of 27 (43%) studies used a dendrogram to choose the number of clusters to include in their study (Table 6). Note that 18 of the 22 studies that applied k-means clustering used hierarchical cluster as a first step to identify the likely number of clusters. Of these 18 studies, 11 explicitly stated that the dendrogram was used to choose the number of clusters.

Of the 8 (13%) studies that specified a maximum number of clusters, the maximum number ranged between 2 and 15 clusters. Seven (11%) studies used a statistic (or multiple statistics), including the c-index [60], Gap statistic [37], deviation from ideal stability [30], Calinski and Harabasz index [30], Dunn's partition [57], cubic cluster criterion (CCC) statistic [28], pseudo-F statistic [28,36], and pseudo-T2 statistic [28,36].

Four studies (6%) avoided very small clusters. Approaches to this include merging 2 clusters containing 6 and 12 samples [61], omitting small clusters containing 1 [35] and 6 [62] samples, and choosing the number such that no cluster contained less than 10% of the total samples [63].

Stability

A total of 11 (17%) studies tested the stability of their cluster solution; the methods are detailed in Table 6. Of these, 1 study used 2 methods, and the remaining 10 each used only 1 method to test stability.

Quality

A total of 24 (38%) studies assessed the quality of their solution using methods beyond those assessing stability. The methods are detailed in Table 6. Of these, 3 used more than one method. The maximum number of methods used in this study was 4.

Of the 30 studies that assessed the stability or quality of their cluster analysis, 21 (70%) reported their findings. However, the reporting of these results was in many cases brief, consisting of statements such as "the clusters were shown to be stable" without providing supporting evidence.

Table 6. Postprocessing methods used in the asthma studies included in this review.

Method	Values, n (%)[a]
Choosing the number of clusters	
Dendrogram	27 (43)
Hierarchical clustering with Ward linkage	19 (30)
Specify a maximum number of clusters[b]	8 (13)
Statistic(s)	7 (11)
Silhouette plot or average silhouette width	5 (8)
Bayesian information criterion	4 (6)
Specify a minimum size of smallest cluster[b]	4 (6)
Previous studies	3 (5)
Unclear	3 (5)
Clinical interpretation	2 (3)
Scree plot	1 (2)
Stability	
Repeated in random subset	3 (5)
Leave-one-out cross-validation	3 (5)
Bootstrap methods	3 (5)
Unclear methods	2 (3)
Train and test set	1 (2)
Quality	
Repeated in selected subset	8 (13)
Repeated with difference methods	6 (10)
Repeated with different initial configurations	5 (8)
Repeated in separate cohort	4 (6)
Repeated with altered features	3 (5)
Repeated at different time point	3 (5)
Repeated with different software	1 (2)

[a]Studies may have used more than 1 method.

[b]These methods were not included when calculating the number of methods used to choose the number of clusters.

Discussion

Principal Findings

We identified 63 studies that applied cluster analysis to multimodal clinical data to identify subtypes of asthma. We explored the clustering methodologies and their limitations in detail. The principal finding of this review was that the majority of the reviewed studies have flaws in the application of cluster analysis. Although some of these flaws were related to the multimodal nature of the clinical data, they extended to aspects of cluster analysis, which are agnostic of data type, such as sample size, stability, and reporting of the results.

These findings build on a previous review, which identified limitations such as lack of robustness in feature selection and neglect to specify distance measures in studies using cluster analysis to contribute to our understanding of the spectrum of asthma syndrome [11]. Our review investigated the methods of feature engineering more generally and identified not only neglect to specify dissimilarity measures but also instances in which the dissimilarity measure was inappropriate for the data to which it was applied. In addition, we identified issues related to sample size, cluster analysis methods, choosing the number of clusters, and testing the stability and quality of results. These issues are discussed in the following paragraphs.

A widespread limitation in the reviewed studies was the small sample size. Studies had overall sample sizes as small as 40 patients, with clusters as small as 6 patients. We argue that there is limited utility in clustering data with such small sample sizes: they may result in clusters that are unstable [64] and may therefore lead to selecting fewer clusters than are present in the underlying population from which the data are sampled.

In the following paragraphs, we discussed the limitations of 3 of the feature selection approaches applied by the reviewed

studies. The first approach was to avoid collinearity or multicollinearity or excluding features that were considered to be *clinically redundant*. Although one should avoid including features that are *redundant* (can be completely deduced from a combination of the other cluster features), this is rarely the case. Therefore, removing features inevitably leads to loss of information. We suggest that the removal of features based on redundancy needs to be carefully considered, for example, 2 or more features (some of which may appear univariately redundant) may jointly contribute toward determining a cluster (or similarly toward the estimation of a clinical outcome in a standard supervised learning setup).

The second was the use of PCA or factor analysis to select features, which has a similar motivation to the concept described earlier for discarding statistically correlated features. There are methodological justifications for the use of PCA, factor analysis, or other nonlinear embedding methods for feature transformation [19]. They aim to jointly combine the original features and project them in a new feature space, which may have some useful properties, including interpretation, determining latent feature structure, and improving the clustering or statistical mapping outcomes [16]. However, we suggest exercising caution toward using these methods for feature selection as described in some of the studies summarized in the Results section of this review because they were fundamentally developed toward different aims. Haldar et al used PCA for feature selection in the first publication to apply cluster analysis to identify asthma subtypes [14]. It is possible that other studies used this as a point of reference for these methods, leading to the common application of these methods in the field of asthma subtyping.

The third approach to feature selection was the use of statistical hypothesis tests with respect to outcomes of interest, as done in 2 studies [42,43]. Methods in which an outcome of interest is used to guide feature selection in cluster analysis have been described previously [65,66]. Although these approaches may be useful for situations in which there exists an outcome of particular interest to the clustering problem, the user should be aware of and acknowledge the assumptions made in the process. In the context of the 2 reviewed studies that used this approach, Sakagami et al did not acknowledge the linearity assumption in linear regression [42], whereas Seino et al's method does not account for potentially highly correlated features [43], a concept that is key in feature selection for cluster analysis.

Feature transformation was applied in only 6 studies, and the methods were generally poorly reported. As with cluster analysis, feature encoding and scaling are important considerations in feature transformation, but none of the studies gave adequate details in their methods. The results of feature transformation were also poorly reported. Although the key reason for applying feature transformation methods is to reduce the dimensionality of the dataset, only 2 [31,45] of the 6 studies provided details on the number of features retained. We suggest that the results of PCA, factor analysis, or MCA should include a table of component (or factor) loadings, the number of features retained, and the proportion of variance accounted for in the transformed features.

Most studies explicitly stated the clustering method that they used but were less explicit regarding the preprocessing steps and choice of dissimilarity measure. Hastie et al [16] state, "Specifying an appropriate dissimilarity measure is far more important in obtaining success with clustering than choice of clustering algorithm."

We expand on this statement, further adding that preprocessing steps such as feature scaling and feature encoding are also more important in obtaining success than the choice clustering algorithm. This is in line with the conclusions of Prosperi et al, who demonstrated that clustering using different feature sets and encodings in asthma datasets can lead to different cluster solutions [29]. Both preprocessing steps and dissimilarity measures, along with their relation to clustering algorithms, have been given poor consideration in clustering applications in asthma, as discussed in the following 3 paragraphs.

First, the Euclidean distance was used with mixed-type data in over half of the studies (54%). Although the Euclidean distance is intended for use with continuous data, problems associated with applying it to mixed-type data may be mitigated by carefully considering feature scaling and feature encoding. However, in our review, we found that many studies did not specify their methods for rescaling, and many studies included ordinal and nominal categorical features but did not specify how these would be treated when calculating the Euclidean distances. The lack of consideration of feature scaling and encoding in these cases may have resulted in assigning an unintended weight structure to the cluster features.

Second, 4 studies used Gower coefficient in hierarchical clustering with Ward linkage [36,67-69], and 1 used tree-based distances [49,70]. These studies should be given some credit for using dissimilarities that can handle mixed-type data. However, the application of hierarchical clustering with Ward linkage relies on the properties of the Euclidean distance in the computations. These properties do not hold for Gower coefficient, and hence, errors are perpetuated at each level of the hierarchy. An example that demonstrates this issue is given in Multimedia Appendix 5.

A final point in the use of k-means and hierarchical clustering using the Ward method with mixed cluster features is that the theory underpinning these methods involves the calculation of cluster means. The mean is not an appropriate summary statistic for categorical features, which are more typically summarized by the mode. For this reason, we suggest that k-medoids may be a more appropriate method for mixed-type features used in clustering. Instead of computing each cluster's mean (as with hierarchical clustering using Ward's method and k-means), k-medoids compute each cluster's medoid, defined as the sample in the cluster for which the average dissimilarity to all other samples in the cluster is minimized [55]. In addition, k-medoids do not rely on the properties of the Euclidean distance in the computations, thus avoiding the issue described in the previous paragraph. Despite these advantages, only 2 studies in this review used k-medoids [30,56].

The SPSS TwoStep method was used in 7 of the 63 studies investigated here. We see 2 key limitations with the application of this method across the reviewed studies. First, none of the

studies gave adequate consideration to the distributional assumptions made when using the log-likelihood distance, and most did not mention the assumptions at all. Second, this method is designed for clustering several millions of samples with many features within an acceptable time and makes a key compromise in doing so [52]. This compromise is that the data are not stored in the main memory but are read sequentially, hence making the solution sensitive to the ordering of the data. None of the studies acknowledged this inherent shortcoming, nor did they confirm that their data were in a random order. Perhaps, more concerningly, the studies that applied these methods actually had very small datasets (range 84-349 samples) that could easily be stored, therefore making other standard techniques more appropriate. In our view, this compromise was therefore unnecessary.

Only 1 study [57] used a method that obtains a *fuzzy* cluster solution (in which a patient may be assigned a membership value to multiple clusters), as opposed to a *hard* cluster solution (in which each patient is assigned to a single cluster) [23]. A fuzzy cluster solution can indicate where a patient membership value is similar across multiple clusters, whereas this information is lost (or leads to lack of stability) in a hard cluster solution. Owing to the noisy nature of clinical data and the clinical complexity of grouping patients into distinct groups, we suggest that fuzzy cluster solutions may be more appropriate than hard cluster solutions in the review applications in asthma. However, it is important to acknowledge that there are added challenges in the interpretation and communication of fuzzy cluster solutions and that the methods may be more computationally intensive [71].

Selecting the number of clusters can be challenging and depends largely on the context of the application. In the case of the reviewed applications in asthma, the *true* number of clusters is unknown, and the analyses are exploratory. Although 86% of the review studies gave some details regarding their methods for choosing the number of clusters (k), they were generally poorly reported. The most popular approach was the dendrogram, but only Labor et al [72] specified their criteria for cutting the dendrogram. In 14 studies, the dendrogram was the only method mentioned. We suggest that more than one method should be used to select the number of clusters to validate this decision.

Our review shows that studies rarely tested the stability and quality of their results, with a particular lack of emphasis on stability. This is concerning, as many studies use methods such as k-means, which reach local minima, and apply them to small sample sizes, thus increasing the risk of obtaining unstable results. We argue that because of the unsupervised nature of cluster analysis, testing the stability and quality of the results should be a key theme and would like to urge researchers and peer reviewers in this research field to carefully consider these aspects. However, we do appreciate that assessing the stability and quality of a solution in the absence of *ground truth* is challenging and that there are currently no well-established frameworks for doing so [27].

Although this review focused on applications in subtyping asthma, the identified issues have been found in studies using cluster analysis to subtype other diseases. For example, recent studies in autism [73] and hypersomnolence [74] have applied cluster analysis to very small samples (55 and 17 patients, respectively). A recent study on Parkinson disease [75] stated in the main text that a *model-based* cluster analysis method was used, whereas the supplementary materials revealed that the method was in fact k-means, which is not model-based. In addition, supplementary materials listed 3 methods for choosing the number of clusters (CCC, pseudo-F, and R-squared statistics) but did not present the results from these 3 methods anywhere in the main text or supplementary materials. These findings demonstrate the widespread nature of the issues that this review has highlighted, and that the issues are not restricted to asthma-related studies.

For a recent example of a well-considered and well-reported application of cluster analysis to multimodal clinical data, we refer the reader to Pikoula et al's study of Chronic Obstructive Pulmonary Disease subtypes [76]. The main text and supplementary materials provide a transparent report of the methodology with respect to feature engineering and cluster analysis methods. In particular, Pikoula et al performed a rigorous assessment of the stability, reproducibility, and sensitivity of the resulting clusters, which could be used as a framework for future studies. The results that were key to the study's conclusions (eg, MCA feature loadings, silhouette plots, results from stability, reproducibility, and sensitivity analyses) are correctly reported in the manuscript, enabling readers to have a thorough understanding of the study's findings.

Limitations

The literature search presented in this study is comprehensive but practically cannot be exhaustive. We restricted the search to articles that included the terms *cluster analysis* or *clustering**. Although it is not strictly speaking correct to do so, some studies in the medical literature use the term *classification* to refer to cluster analysis, often confusing the 2 terms and sometimes using them almost interchangeably, for example, see the studies by Just et al [34] and Kim et al [46]. Widening the search to identify studies that use the term *classification* would have greatly increased the initial number of results of the PubMed search, but we suspect that the increase in the number of eligible studies for cluster analysis identified would have been small. Similarly, the terms *latent class analysis* and *mixture model analysis* might sometimes be erroneously used to refer to cluster analysis: we clarify that these terms were not included in our search strategy. As this is not a systematic review, we feel that our search criteria are fully sufficient for this study's purposes.

We did not fully explore multiple kernel k-means [77] or spectral clustering [78] methods, each used by 1 study in this review. As with all other cluster analysis methods mentioned here, careful consideration must be taken when applying these methods to mixed-type data. There are numerous other considerations that are important to these methods, such as the choice of kernel function, but these are beyond the scope of this review.

Conclusions

This review highlights a number of issues in previous applications of cluster analysis to multimodal clinical data in asthma. We make the following key recommendations based on these findings:

- Careful consideration should be given to the preprocessing of multimodal clinical data and how the scaling and encoding of features may affect their weighting in the analysis.
- The choice of dissimilarity measures and cluster analysis methods are dependent on one another as well as on the scaling and encoding of the data. Certain combinations of these data analytics components may be incompatible and give unreliable results.
- The stability and quality of the cluster results should be thoroughly evaluated.

The abovementioned recommendations focus on the application of cluster analysis, but we put similar emphasis on the clear reporting of each of the abovementioned points, as this was also found to be lacking in the reviewed papers.

Acknowledgments

This study was supported by the Health Data Research, United Kingdom (HDR UK), which receives funding from HDR UK Ltd (HDR-5012) funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), the British Heart Foundation, and the Wellcome Trust and by the Asthma UK Centre for Applied Research, which is funded by Asthma UK. The funders had no role in the study or the decision to submit this work to be considered for publication.

Authors' Contributions

EH was responsible for conducting the study. EH conducted the identification of articles and screened them for eligibility. EH and HT independently extracted data according to the described methodology and synthesized the findings. EH wrote up the first draft of the manuscript, and AT, AS, and HT contributed to the final version.

Conflicts of Interest

AS is supported by a research grant from the Asthma UK Centre for Applied Research. All other authors have no conflict of interest pertaining to this study to declare.

Multimedia Appendix 1
Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist.
[DOC File , 64 KB - medinform_v8i5e16452_app1.doc]

Multimedia Appendix 2
Data dictionary.
[DOCX File , 22 KB - medinform_v8i5e16452_app2.docx]

Multimedia Appendix 3
Study characteristics.
[DOCX File , 85 KB - medinform_v8i5e16452_app3.docx]

Multimedia Appendix 4
Breakdown of methods used by the 11 studies that did not use the three most common clustering methods.
[DOCX File , 16 KB - medinform_v8i5e16452_app4.docx]

Multimedia Appendix 5
Illustrative example of the use of Gower coefficient with hierarchical clustering and Ward linkage.
[DOCX File , 12 KB - medinform_v8i5e16452_app5.docx]

References

1. Lawton M, Ben-Shlomo Y, May MT, Baig F, Barber TR, Klein JC, et al. Developing and validating Parkinson's disease subtypes and their motor and cognitive progression. J Neurol Neurosurg Psychiatry 2018 Dec;89(12):1279-1287 [FREE Full text] [doi: 10.1136/jnnp-2018-318337] [Medline: 30464029]

2. Ousley O, Cermak T. Autism spectrum disorder: defining dimensions and subgroups. Curr Dev Disord Rep 2014 Mar 1;1(1):20-28 [FREE Full text] [doi: 10.1007/s40474-013-0003-1] [Medline: 25072016]
3. Li D, Haritunians T, Landers C, Potdar AA, Yang S, Huang H, et al. Late-onset Crohn's disease is a subgroup distinct in genetic and behavioral risk factors with UC-like characteristics. Inflamm Bowel Dis 2018 Oct 12;24(11):2413-2422 [FREE Full text] [doi: 10.1093/ibd/izy148] [Medline: 29860388]
4. Bowman P, Flanagan SE, Hattersley AT. Future roadmaps for precision medicine applied to diabetes: rising to the challenge of heterogeneity. J Diabetes Res 2018;2018:3061620 [FREE Full text] [doi: 10.1155/2018/3061620] [Medline: 30599002]
5. Sidhaye VK, Nishida K, Martinez FJ. Precision medicine in COPD: where are we and where do we need to go? Eur Respir Rev 2018 Sep 30;27(149) [FREE Full text] [doi: 10.1183/16000617.0022-2018] [Medline: 30068688]
6. Zhang J, Späth SS, Marjani SL, Zhang W, Pan X. Characterization of cancer genomic heterogeneity by next-generation sequencing advances precision medicine in cancer treatment. Precis Clin Med 2018 Jun;1(1):29-48 [FREE Full text] [doi: 10.1093/pcmedi/pby007] [Medline: 30687561]
7. Pavord ID, Beasley R, Agusti A, Anderson GP, Bel E, Brusselle G, et al. After asthma: redefining airways diseases. Lancet 2018 Jan 27;391(10118):350-400. [doi: 10.1016/S0140-6736(17)30879-6] [Medline: 28911920]
8. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1789-1858 [FREE Full text] [doi: 10.1016/S0140-6736(18)32279-7] [Medline: 30496104]
9. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980-2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2018 Nov 10;392(10159):1736-1788 [FREE Full text] [doi: 10.1016/S0140-6736(18)32203-7] [Medline: 30496103]
10. Rackemann FM. A working classification of asthma. Am J Med 1947 Nov;3(5):601-606. [doi: 10.1016/0002-9343(47)90204-0] [Medline: 20269240]
11. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. Pulm Ther 2016;2:19-41 [FREE Full text] [doi: 10.1007/s41030-016-0017-z] [Medline: 27512723]
12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, USA: Springer; 2009.
13. Hennig C, Liao TF. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. J R Stat Soc Ser C Appl Stat 2013;62(3):309-369. [doi: 10.1111/j.1467-9876.2012.01066.x]
14. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med 2008 Aug 1;178(3):218-224 [FREE Full text] [doi: 10.1164/rccm.200711-1754OC] [Medline: 18480428]
15. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. Curr Allergy Asthma Rep 2015 Jul;15(7):38 [FREE Full text] [doi: 10.1007/s11882-015-0542-0] [Medline: 26143394]
16. Hastie T, Tibshirani R, Friedman J. Unsupervised learning. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2009:485-552.
17. Dolnicar S, Grün B, Leisch F, Schmidt K. Required sample sizes for data-driven market segmentation analyses in tourism. J Travel Res 2014;53(3):296-306. [doi: 10.1177/0047287513496475]
18. Hastie T, Tibshirani R, Friedman J. Overview of supervised learning. In: Hastie T, Tibshirani R, Friedman J, editors. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (chapter 2). New York: Springer; 2009:9-38.
19. Ben-Hur A, Guyon I. Detecting stable clusters using principal component analysis. Methods Mol Biol 2003;224:159-182. [doi: 10.1385/1-59259-364-X:159] [Medline: 12710673]
20. van der Maaten L, Postma E, van den Herik J. Dimensionality reduction: a comparative review. J Mach Learn Res 2009;10:66-71 [FREE Full text]
21. Jackson E. A User's Guide to Principal Components. Jersey City, USA: Wiley-Blackwell; 1991.
22. Pagès J. Multiple correspondence analysis. In: Multiple factor Analysis by Example using R. Boca Raton, Florida: Chapman and Hall/CRC; 2018:39-66.
23. Pagès J. Multiple factor analysis and procrustes analysis. In: Multiple factor Analysis by Example using R. Boca Raton, Florida: Chapman and Hall/CRC; 2018:189-208.
24. Gower JC. A general coefficient of similarity and some of its properties. Biometrics 1971 Dec;27(4):857-871. [doi: 10.2307/2528823]
25. Everitt B, Landau S, Leese M. Cluster Analysis. Fifth Edition. New York, USA: Wiley Publishing; 2011.
26. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987 Nov;20(5):53-65. [doi: 10.1016/0377-0427(87)90125-7]
27. von Luxburg U. Clustering stability: an overview. Found Trends Mach Learn 2010;2(3):235-274. [doi: 10.1561/2200000008]
28. Schatz M, Hsu JY, Zeiger RS, Chen W, Dorenbaum A, Chipps BE, et al. Phenotypes determined by cluster analysis in severe or difficult-to-treat asthma. J Allergy Clin Immunol 2014 Jun;133(6):1549-1556. [doi: 10.1016/j.jaci.2013.10.006] [Medline: 24315502]

29. Prosperi MC, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, Simpson A, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med 2013 Dec 1;188(11):1303-1312 [FREE Full text] [doi: 10.1164/rccm.201304-0694OC] [Medline: 24180417]
30. Lefaudeux D, de Meulder B, Loza MJ, Peffer N, Rowe A, Baribaud F, U-BIOPRED Study Group. U-BIOPRED clinical adult asthma clusters linked to a subset of sputum omics. J Allergy Clin Immunol 2017 Jun;139(6):1797-1807. [doi: 10.1016/j.jaci.2016.08.048] [Medline: 27773852]
31. Khusial RJ, Sont JK, Loijmans RJ, Snoeck-Stroband JB, Assendelft PJ, Schermer TR, ACCURATE Study Group. Longitudinal outcomes of different asthma phenotypes in primary care, an observational study. NPJ Prim Care Respir Med 2017 Oct 3;27(1):55 [FREE Full text] [doi: 10.1038/s41533-017-0057-3] [Medline: 28974677]
32. Hsiao H, Lin M, Wu C, Wang C, Wang T. Sex-specific asthma phenotypes, inflammatory patterns, and asthma control in a cluster analysis. J Allergy Clin Immunol Pract 2019 Feb;7(2):556-67.e15. [doi: 10.1016/j.jaip.2018.08.008] [Medline: 30170162]
33. Moore WC, Hastie AT, Li X, Li H, Busse WW, Jarjour NN, National Heart, Lung, Blood Institute's Severe Asthma Research Program. Sputum neutrophil counts are associated with more severe asthma phenotypes using cluster analysis. J Allergy Clin Immunol 2014 Jun;133(6):1557-63.e5 [FREE Full text] [doi: 10.1016/j.jaci.2013.10.011] [Medline: 24332216]
34. Just J, Gouvis-Echraghi R, Rouve S, Wanin S, Moreau D, Annesi-Maesano I. Two novel, severe asthma phenotypes identified during childhood using a clustering approach. Eur Respir J 2012 Jul;40(1):55-60 [FREE Full text] [doi: 10.1183/09031936.00123411] [Medline: 22267763]
35. Pérez-Losada M, Authelet KJ, Hoptay CE, Kwak C, Crandall KA, Freishtat RJ. Pediatric asthma comprises different phenotypic clusters with unique nasal microbiotas. Microbiome 2018 Oct 4;6(1):179 [FREE Full text] [doi: 10.1186/s40168-018-0564-7] [Medline: 30286807]
36. Ding L, Li D, Wathen M, Altaye M, Mersha TB. African ancestry is associated with cluster-based childhood asthma subphenotypes. BMC Med Genomics 2018 May 31;11(1):51 [FREE Full text] [doi: 10.1186/s12920-018-0367-5] [Medline: 29855310]
37. Howrylak JA, Fuhlbrigge AL, Strunk RC, Zeiger RS, Weiss ST, Raby BA, Childhood Asthma Management Program Research Group. Classification of childhood asthma phenotypes and long-term clinical responses to inhaled anti-inflammatory medications. J Allergy Clin Immunol 2014 May;133(5):1289-300, 1300.e1 [FREE Full text] [doi: 10.1016/j.jaci.2014.02.006] [Medline: 24892144]
38. Loureiro CC, Sa-Couto P, Todo-Bom A, Bousquet J. Cluster analysis in phenotyping a Portuguese population. Rev Port Pneumol (2006) 2015 Sep 3 [Online ahead of print]. [doi: 10.1016/j.rppnen.2015.07.006] [Medline: 26344641]
39. Amore M, Antonucci C, Bettini E, Boracchia L, Innamorati M, Montali A, et al. Disease control in patients with asthma is associated with alexithymia but not with depression or anxiety. Behav Med 2013;39(4):138-145. [doi: 10.1080/08964289.2013.818931] [Medline: 24236811]
40. Cabral AL, Sousa AW, Mendes FA, Carvalho CR. Phenotypes of asthma in low-income children and adolescents: cluster analysis. J Bras Pneumol 2017;43(1):44-50 [FREE Full text] [doi: 10.1590/S1806-37562016000000039] [Medline: 28125150]
41. Wang L, Liang R, Zhou T, Zheng J, Liang BM, Zhang HP, et al. Identification and validation of asthma phenotypes in Chinese population using cluster analysis. Ann Allergy Asthma Immunol 2017 Oct;119(4):324-332. [doi: 10.1016/j.anai.2017.07.016] [Medline: 28866310]
42. Sakagami T, Hasegawa T, Koya T, Furukawa T, Kawakami H, Kimura Y, et al. Cluster analysis identifies characteristic phenotypes of asthma with accelerated lung function decline. J Asthma 2014 Mar;51(2):113-118. [doi: 10.3109/02770903.2013.852201] [Medline: 24102534]
43. Seino Y, Hasegawa T, Koya T, Sakagami T, Mashima I, Shimizu N, Niigata Respiratory Disease Study Group. A cluster analysis of bronchial asthma patients with depressive symptoms. Intern Med 2018 Jul 15;57(14):1967-1975 [FREE Full text] [doi: 10.2169/internalmedicine.9073-17] [Medline: 29526967]
44. Agache I, Strasser DS, Klenk A, Agache C, Farine H, Ciobanu C, et al. Serum IL-5 and IL-13 consistently serve as the best predictors for the blood eosinophilia phenotype in adult asthmatics. Allergy 2016 Aug;71(8):1192-1202. [doi: 10.1111/all.12906] [Medline: 27060452]
45. Newby C, Heaney LG, Menzies-Gow A, Niven RM, Mansur A, Bucknall C, British Thoracic Society Severe Refractory Asthma Network. Statistical cluster analysis of the British Thoracic Society Severe refractory Asthma Registry: clinical outcomes and phenotype stability. PLoS One 2014;9(7):e102987 [FREE Full text] [doi: 10.1371/journal.pone.0102987] [Medline: 25058007]
46. Kim MA, Shin SW, Park JS, Uh ST, Chang HS, Bae DJ, et al. Clinical characteristics of exacerbation-prone adult asthmatics identified by cluster analysis. Allergy Asthma Immunol Res 2017 Nov;9(6):483-490 [FREE Full text] [doi: 10.4168/aair.2017.9.6.483] [Medline: 28913987]
47. Sendín-Hernández MP, Ávila-Zarza C, Sanz C, García-Sánchez A, Marcos-Vadillo E, Muñoz-Bellido FJ, et al. Cluster analysis identifies 3 phenotypes within allergic asthma. J Allergy Clin Immunol Pract 2018;6(3):955-61.e1. [doi: 10.1016/j.jaip.2017.10.006] [Medline: 29133218]
48. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963 Mar;58(301):236-244. [doi: 10.1080/01621459.1963.10500845]

49. Buttrey SE, Whitaker LR. treeClust: an R package for tree-based clustering dissimilarities. R J 2015;7(2):227-236 [FREE Full text] [doi: 10.32614/rj-2015-032]
50. Meyer N, Nuss SJ, Rothe T, Siebenhüner A, Akdis CA, Menz G. Differential serum protein markers and the clinical severity of asthma. J Asthma Allergy 2014;7:67-75 [FREE Full text] [doi: 10.2147/JAA.S53920] [Medline: 24851055]
51. Zhang T, Ramakrishnan R, Livny M. BIRCH: a new data clustering algorithm and its applications. Data Min Knowl Discov 1996;25(2):141-182. [doi: 10.1145/233269.233324]
52. Bacher J, Wenzig K, Vogler M. SPSS TwoStep Cluster - a first evaluation. Soc Sci Open Access Repos 2004 [FREE Full text]
53. Ruggieri S, Drago G, Longo V, Colombo P, Balzan M, Bilocca D, RESPIRA Project Group. Sensitization to dust mite defines different phenotypes of asthma: a multicenter study. Pediatr Allergy Immunol 2017 Nov;28(7):675-682. [doi: 10.1111/pai.12768] [Medline: 28783215]
54. Ye W, Xu W, Guo X, Han F, Peng J, Li X, et al. Differences in airway remodeling and airway inflammation among moderate-severe asthma clinical phenotypes. J Thorac Dis 2017 Sep;9(9):2904-2914 [FREE Full text] [doi: 10.21037/jtd.2017.08.01] [Medline: 29221262]
55. Kaufman L, Rousseeuw PJ. Partition Around Medoids (Program PAM). In: Kaufman L, Rousseeuw PJ, editors. Finding Groups in Data: An Introduction to Cluster Analysis. NJ: John Wiley & Sons; 2005.
56. Sekiya K, Nakatani E, Fukutomi Y, Kaneda H, Iikura M, Yoshida M, et al. Severe or life-threatening asthma exacerbation: patient heterogeneity identified by cluster analysis. Clin Exp Allergy 2016 Aug;46(8):1043-1055. [doi: 10.1111/cea.12738] [Medline: 27041475]
57. Loza MJ, Djukanovic R, Chung KF, Horowitz D, Ma K, Branigan P, ADEPT (Airways Disease Endotyping for Personalized Therapeutics), U-BIOPRED (Unbiased Biomarkers for the Prediction of Respiratory Disease Outcome Consortium) investigators. Validated and longitudinally stable asthma phenotypes based on cluster analysis of the ADEPT study. Respir Res 2016 Dec 15;17(1):165 [FREE Full text] [doi: 10.1186/s12931-016-0482-9] [Medline: 27978840]
58. Dhillon IS, Guan Y, Kulis B. Kernel K-Means: Spectral Clustering and Normalized Cuts. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. 2004 Presented at: KDD'04; August 22 - 25, 2004; Seattle, WA, USA p. 551-556. [doi: 10.1145/1014052.1014118]
59. Wu W, Bang S, Bleecker ER, Castro M, Denlinger L, Erzurum SC, et al. Multiview cluster analysis identifies variable corticosteroid response phenotypes in severe asthma. Am J Respir Crit Care Med 2019 Jun 1;199(11):1358-1367. [doi: 10.1164/rccm.201808-1543OC] [Medline: 30682261]
60. Gomez JL, Yan X, Holm CT, Grant N, Liu Q, Cohn L, SARP Investigators. Characterisation of asthma subgroups associated with circulating YKL-40 levels. Eur Respir J 2017 Oct;50(4) [FREE Full text] [doi: 10.1183/13993003.00800-2017] [Medline: 29025889]
61. Amelink M, de Nijs SB, de Groot JC, van Tilburg PM, van Spiegel PI, Krouwels FH, et al. Three phenotypes of adult-onset asthma. Allergy 2013;68(5):674-680. [doi: 10.1111/all.12136] [Medline: 23590217]
62. Benton AS, Wang Z, Lerner J, Foerster M, Teach SJ, Freishtat RJ. Overcoming heterogeneity in pediatric asthma: tobacco smoke and asthma characteristics within phenotypic clusters in an African American cohort. J Asthma 2010 Sep;47(7):728-734 [FREE Full text] [doi: 10.3109/02770903.2010.491142] [Medline: 20684733]
63. Lemiere C, NGuyen S, Sava F, D'Alpaos V, Huaux F, Vandenplas O. Occupational asthma phenotypes identified by increased fractional exhaled nitric oxide after exposure to causal agents. J Allergy Clin Immunol 2014 Nov;134(5):1063-1067. [doi: 10.1016/j.jaci.2014.08.017] [Medline: 25262466]
64. Garge NR, Page GP, Sprague AP, Gorman BS, Allison DB. Reproducible clusters from microarray research: whither? BMC Bioinformatics 2005 Jul 15;6(Suppl 2):S10 [FREE Full text] [doi: 10.1186/1471-2105-6-S2-S10] [Medline: 16026595]
65. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004 Apr;2(4):E108 [FREE Full text] [doi: 10.1371/journal.pbio.0020108] [Medline: 15094809]
66. Bair E. Semi-supervised clustering methods. Wiley Interdiscip Rev Comput Stat 2013;5(5):349-361 [FREE Full text] [doi: 10.1002/wics.1270] [Medline: 24729830]
67. Just J, Gouvis-Echraghi R, Couderc R, Guillemot-Lambert N, Saint-Pierre P. Novel severe wheezy young children phenotypes: boys atopic multiple-trigger and girls nonatopic uncontrolled wheeze. J Allergy Clin Immunol 2012 Jul;130(1):103-10.e8. [doi: 10.1016/j.jaci.2012.02.041] [Medline: 22502798]
68. Just J, Saint-Pierre P, Gouvis-Echraghi R, Boutin B, Panayotopoulos V, Chebahi N, et al. Wheeze phenotypes in young children have different courses during the preschool period. Ann Allergy Asthma Immunol 2013 Oct;111(4):256-61.e1. [doi: 10.1016/j.anai.2013.07.002] [Medline: 24054360]
69. Just J, Saint-Pierre P, Gouvis-Echraghi R, Laoudi Y, Roufai L, Momas I, et al. Childhood allergic asthma is not a single phenotype. J Pediatr 2014 Apr;164(4):815-820. [doi: 10.1016/j.jpeds.2013.11.037] [Medline: 24412137]
70. Zoratti EM, Krouse RZ, Babineau DC, Pongracic JA, O'Connor GT, Wood RA, et al. Asthma phenotypes in inner-city children. J Allergy Clin Immunol 2016 Oct;138(4):1016-1029 [FREE Full text] [doi: 10.1016/j.jaci.2016.06.061] [Medline: 27720016]
71. Kaufman L, Rousseeuw PJ. Fuzzy Analysis (Program FANNY). In: Kaufman L, Rousseeuw PJ, editors. Finding Groups in Data: An Introduction to Cluster Analysis. NJ: John Wiley & Sons; 2005:164-198.

72. Labor M, Labor S, Jurić I, Fijačko V, Grle SP, Plavec D. Mood disorders in adult asthma phenotypes. J Asthma 2018 Jan;55(1):57-65. [doi: 10.1080/02770903.2017.1306546] [Medline: 28489959]
73. Obara T, Ishikuro M, Tamiya G, Ueki M, Yamanaka C, Mizuno S, et al. Potential identification of vitamin B6 responsiveness in autism spectrum disorder utilizing phenotype variables and machine learning methods. Sci Rep 2018 Oct 4;8(1):14840 [FREE Full text] [doi: 10.1038/s41598-018-33110-w] [Medline: 30287864]
74. Cook JD, Rumble ME, Plante DT. Identifying subtypes of Hypersomnolence Disorder: a clustering analysis. Sleep Med 2019 Dec;64:71-76. [doi: 10.1016/j.sleep.2019.06.015] [Medline: 31670163]
75. Wolters AF, Moonen AJ, Lopes R, Leentjens AF, Duits AA, Defebvre L, et al. Grey matter abnormalities are associated only with severe cognitive decline in early stages of Parkinson's disease. Cortex 2020 Feb;123:1-11. [doi: 10.1016/j.cortex.2019.09.015] [Medline: 31733342]
76. Pikoula M, Quint JK, Nissen F, Hemingway H, Smeeth L, Denaxas S. Identifying clinically important COPD sub-types using data-driven approaches in primary care population based electronic health records. BMC Med Inform Decis Mak 2019 Apr 18;19(1):86 [FREE Full text] [doi: 10.1186/s12911-019-0805-0] [Medline: 30999919]
77. Bang S, Yu Y, Wu W. Robust multiple kernel k-means clustering using min-max optimization. arXiv preprints 2018 preprint; arXiv:1803.02458 [FREE Full text]
78. Ng A, Jordan M, Weiss Y. On Spectral Clustering: Analysis and an Algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. 2001 Presented at: NIPS'01; December 3-8, 2001; Vancouver, Canada p. 849-856 URL: https://dl.acm.org/doi/10.5555/2980539.2980649

Abbreviations

CCC: cubic cluster criterion
HDR: Health Data Research
MCA: multiple correspondence analysis
PCA: principal component analysis
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 30.09.19; peer-reviewed by M Pikoula, C Newby, K Usop; comments to author 11.11.19; revised version received 10.12.19; accepted 10.02.20; published 28.05.20.

Please cite as:
Horne E, Tibble H, Sheikh A, Tsanas A
Challenges of Clustering Multimodal Clinical Data: Review of Applications in Asthma Subtyping
JMIR Med Inform 2020;8(5):e16452
URL: http://medinform.jmir.org/2020/5/e16452/
doi:10.2196/16452
PMID:32463370

Original Paper

Detecting False Alarms by Analyzing Alarm-Context Information: Algorithm Development and Validation

Chrystinne Fernandes[1*], PhD; Simon Miles[2*], PhD; Carlos José Pereira Lucena[1*], PhD

[1]Department of Informatics, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil

[2]Department of Informatics, King's College London, London, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Chrystinne Fernandes, PhD
Department of Informatics
Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
RDC Bldg, 4th Fl
225 Marquês de São Vicente St
Rio de Janeiro, 22451-900
Brazil
Phone: 55 21 3527 1510
Email: chrystinne@gmail.com

Abstract

Background: Although alarm safety is a critical issue that needs to be addressed to improve patient care, hospitals have not given serious consideration about how their staff should be using, setting, and responding to clinical alarms. Studies have indicated that 80%-99% of alarms in hospital units are false or clinically insignificant and do not represent real danger for patients, leading caregivers to miss relevant alarms that might indicate significant harmful events. The lack of use of any intelligent filter to detect recurrent, irrelevant, and/or false alarms before alerting health providers can culminate in a complex and overwhelming scenario of sensory overload for the medical team, known as *alarm fatigue*.

Objective: This paper's main goal is to propose a solution to mitigate *alarm fatigue* by using an automatic reasoning mechanism to decide how to calculate false alarm probability (FAP) for alarms and whether to include an indication of the FAP (ie, FAP_LABEL) with a notification to be visualized by health care team members designed to help them prioritize which alerts they should respond to next.

Methods: We present a new approach to cope with the *alarm fatigue* problem that uses an automatic reasoner to decide how to notify caregivers with an indication of FAP. Our reasoning algorithm calculates FAP for alerts triggered by sensors and multiparametric monitors based on statistical analysis of false alarm indicators (FAIs) in a simulated environment of an intensive care unit (ICU), where a large number of warnings can lead to *alarm fatigue*.

Results: The main contributions described are as follows: (1) a list of FAIs we defined that can be utilized and possibly extended by other researchers, (2) a novel approach to assess the probability of a false alarm using statistical analysis of multiple inputs representing alarm-context information, and (3) a reasoning algorithm that uses alarm-context information to detect false alarms in order to decide whether to notify caregivers with an indication of FAP (ie, FAP_LABEL) to avoid *alarm fatigue*.

Conclusions: Experiments were conducted to demonstrate that by providing an intelligent notification system, we could decide how to identify false alarms by analyzing alarm-context information. The reasoner entity we described in this paper was able to attribute FAP values to alarms based on FAIs and to notify caregivers with a FAP_LABEL indication without compromising patient safety.

(JMIR Med Inform 2020;8(5):e15407) doi:10.2196/15407

KEYWORDS

alarm fatigue; alarm safety; false alarms; eHealth systems; remote patient monitoring; notification; reasoning; sensors

Introduction

Overview

In our previous work [1], we developed a software framework for remote patient monitoring with notification capabilities that were handled by the use of software agents. In the systems built through our framework, the anomaly detection process worked by triggering an alarm every time an anomaly occurred, independent of the circumstances [2,3].

However, these alerts are often false alarms that do not represent real danger for patients. In this case, the lack of use of any intelligent filter to detect an indication of false alarms before alerting health providers can culminate in a context of a sensory overload for the medical team. This context can result in alarm fatigue and compromise health providers' attention, leading them to miss relevant alarms that might announce significant harmful events.

As a strategy to mitigate the alarm fatigue issue, we present a new approach to monitor patients by using an intelligent notification process supported by a reasoning mechanism. This mechanism associates a false alarm probability (FAP) to alarms based on their real-time context information, including (1) information about a patient's circumstances, such as his or her repositioning in bed, and localization, which is tracked in real time using wearable devices with GPS, and (2) information about sensors, including battery charge life, the last time the patient's skin was prepared to receive electrodes, and the last time electrodes were changed, among others.

After receiving this context information as input, the reasoner's work begins by analyzing each alarm and calculating the FAP associated with it according to the false alarm indicators (FAIs) we defined, based on our literature review. Thus, the reasoner uses the FAP calculated for each alarm to decide whether to include an indication of false alarm probability (ie, FAP_LABEL) with a notification that can be visualized by caregivers.

This paper's main goal is to propose a solution to mitigate alarm fatigue by using an automatic reasoning mechanism to assist caregivers in their decision-making process of choosing which alarms they should respond to next. Our specific goal is to attribute an FAP to each alert based on the context in which it has been generated, such as a patient's condition and information about monitoring devices and sensors. We aim to determine the probability of an alarm being a false alarm in order to decide whether to include this information (ie, FAP_LABEL) with the notifications sent to caregivers.

We addressed the following research questions: (1) How can an automatic reasoning system calculate an indication of FAP for an alarm generated by sensors and monitoring devices? (2) How can we decide whether to add an FAP_LABEL to a notification that could be visualized by the health care team?

We defined the following hypotheses for our case study:

1. Hypothesis 1 (H1): Our reasoning algorithm should associate an FAP value to every alarm generated by sensors and monitoring devices in our experiments.
2. Hypothesis 2 (H2): Our reasoning algorithm should add an indication of an FAP to each alarm, upon which the reasoner should decide whether or not to notify caregivers with an indication of FAP (ie, FAP_LABEL).
3. Hypothesis 3 (H3): Patient safety should not be compromised if and when the reasoning algorithm decides to add an FAP_LABEL to the notification.

The main contributions of this work are as follows:

1. A list of the FAIs we defined that can be utilized and possibly extended by other researchers.
2. A novel approach to assess the probability of a false alarm using statistical analysis of multiple inputs representing alarm-context information.
3. A reasoning algorithm that uses alarm-context information to detect false alarms in order to decide whether to notify caregivers with an indication of FAP (ie, FAP_LABEL) to avoid alarm fatigue.

Background and Related Work

Alarms and the Impact of Alarm Safety in Patient Care

Alarms are utilized to improve patient safety and quality of care by detecting changes early and requiring appropriate action. However, the medical literature contains many studies showing that up to 90% of all alarms in critical-care monitoring are false positives. The vast majority of all threshold alarms in the intensive care unit (ICU) do not have a real clinical impact on the care of the critically ill [4].

Many studies have recorded the number of alerts being triggered nowadays in ICUs during a period of time in order to analyze the impact of alarm safety in patient care as a consequence of the excessive volume of alarms. For instance, Kierra reported that during a 12-day analysis of the alarm system at The Johns Hopkins Hospital in Baltimore, USA, there was an average of 350 alerts per bed per day and that in one ICU, the average was 771 alerts per bed per day [5].

Lawless analyzed alarm soundings that occurred in an ICU during a 7-day period, recorded by ICU staff [6]. In his experiments, he categorized alarms into three types: false, significant (ie, resulted in change in therapy), or induced (ie, by staff manipulations; not significant). He showed that out of 2176 total alarm soundings, 1481 (68.06%) were false, 119 (5.47%) were significant, and 576 (26.47%) were induced. His results showed that over 94% of alarm soundings in a pediatric ICU may not be clinically important. Based on his findings, the author concluded that current monitoring systems are poor predictors of untoward events.

In addition to the excessive number of alarms in ICUs, another alarm-related problem, as presented by Sendelbach, is the high number of different alarm signals that was reducing the effectiveness of the alarms, creating confusion for staff, and was thus detrimental to patient care [7]. In 1983, up to six alarms could be associated with each patient in an ICU. By 1994, up to 33 different alarms were identified, and by 2011, this number increased to over 40 different alarm signals in an ICU [7]. There have been as many as 120 separate alarm devices in an operating

room (OR) that are stand-alone, uncorrelated, and unprioritized [7].

The main problem of having so many different devices triggering alarms is that it is not feasible for nurses to identify all of them, which means that this increase has occurred despite staff having difficulty in learning all available alarm signals in their work environment. Staff from an OR were only able to identify between 10 and 15 out of the 26 alarms triggered in the room, and ICU nurses could only identify between 9 and 14 out of 23 alarms found in the ICU, which contributes to the alarm overload problem [7].

Kerr and Hayes [8] recognized that the excessive number and many diverse types of alarms were resulting in adverse consequences to patient care, including the following: (1) the reduction of the effectiveness of alarms, (2) creation of confusion and distraction for caregivers, who were having difficulties in responding to alarms, and (3) the deterioration of patient care, putting patients in a more unsafe environment.

Lastly, a third alarm-related problem we are focusing on in this paper is the excessive number of false alarms. Studies have indicated that false and/or clinically insignificant alarms range from 80% to 99% [9]. False alarms are frequently triggered by erroneous or absent patient data. These types of alarms can be caused by events such as patient movement or repositioning in bed and by poor placement of sensors, such as an external fetal heart rate monitor or pulse oximeter [10].

Along with the already-mentioned alarm-related problems that can affect patient care, there is more information in ICUs that is considered critical for the health care team, such as (1) the perceived alarm urgency, and (2) the perceived true alarm rate of the alarm system [10]. Tanner showed that perceived alarm urgency contributes to the nurses' alarm response; however, nurses also use additional strategies to determine response, including the criticality of the patient, signal duration, uncommonness of the alarming device, and workload [10].

Regarding the perceived true alarm rate of the alarm system, an important finding by Tanner is the link between the impact of the perceived true alarm rate of the alarm system by caregivers and its influence on patient care. The author showed that the nurses' responses to alarms follow the perceived true alarm rate of the alarm system. According to the author, if the true alarm rate is perceived to be 10% reliable, then the response rate will be about 10% [10].

Although alarm safety is a critical issue that needs to be addressed to improve patient care, hospitals have not given serious consideration to how their staff should be using, setting, and responding to clinical alarms, according to the Emergency Care Research Institute (ECRI) [11]. Currently, this complex and overwhelming scenario is still a problem that culminated in an unsolved health problem known as *alarm fatigue*, which we next describe.

Alarm Fatigue

By definition, *alarm fatigue* consists of the lack of response due to excessive numbers of alarms in hospital environments, especially in ICUs, resulting in sensory overload and desensitization [9]. This issue has the potential to compromise patient safety [12], since frequent alarms are distracting and interfere with a clinician's performance of critical tasks. Excessive false positive alarms may lead to apathy, resulting in a lower likelihood that real events may be acted on. For their part, insignificant alarms may result in distraction and could lead to the disabling of alarm systems by staff [9].

To illustrate this scenario, studies have indicated that false and/or clinically insignificant alarms range from 80% to 99% [9]. The presence of medical devices generate enough false alarms to cause a reduction in responses, leading to a scenario in which caregivers disable, silence, and/or ignore the alarms [12] or are slow to respond [8,9].

In Table 1, we summarized the information we presented about alarm-related issues as well as their causes, consequences to the staff, consequences to patients' care, and avoidance strategies [9].

Table 1. Summary of alarm-related issues.

Alarm-related issue	Causes	Consequences to the staff	Consequences to patient care	Avoidance strategies
Excessive false positive alarms	Can be attributed to patient manipulation (ie, motion artifact)	Apathy and desensitization Mistrust	Reduction in responding Lack of caregiver response Real events being less likely to be acted on	Suspension of alarms for a short period prior to patient manipulation Statistical methods should be suitable to decrease the number of false positive alarms
Frequent insignificant or irrelevant alarms	Use of the default alarm settings Poor staff education on alarm management	Distraction Reduction in trust	Disruption of patient care Disabling of alarm systems by staff	Eliminating nonessential alarms Adjusting alarm parameters on monitors to suit patients' conditions Staff education on alarm management

Statistical and Artificial Intelligence-Related Approaches

According to Imhoff et al, the quality of medical device alarms is unsatisfactory, affecting quality of care and patient safety. Since the low quality of alarm-generating algorithms is one of the main causes of this problem, major improvements in alarm algorithms are urgently needed [4].

To achieve this goal, a variety of alarm-suppression algorithms have been developed and successfully applied in the laboratory and the clinical environment to avoid alarm fatigue, such as relevance vector machine learning, statistical metrics, time series

analysis, spectral regression, feature selection, and other classifiers [13]. Imhoff et al showed different methods that have been proposed for use in the alarm systems of medical devices, including statistical approaches, such as improved data preprocessing, robust signal extraction, segmentation, median filter, statistical process control, and time series analysis for pattern detection, among others. Artificial intelligence (AI) methods have also been investigated and include approaches based on machine learning, neural networks, random forests, fuzzy logic, and Bayesian networks [4].

Another strategy to avoid alarm fatigue is to use notification delays that are performed through the use of a middleware between the alarming medical device and the clinicians' receiver device, such as a mobile phone or a tablet. Several studies found that introducing alarm delays before notifying caregivers could decrease *false alarms* by 25%-67% [13]. Regarding the reduction of the total alarms, considering the effects of these interventions, alarm quantities decreased between 18.5% and as much as 89%, according to Winters et al. Fernandes et al also present a reasoning algorithm that works through the use of a notification delay strategy to mitigate alarm fatigue [14]. Other examples of promising proposed approaches are the application of contextuality and the integration of alarms to create smart alarms with improved data presentation through human factors engineering [13].

According to Imhoff et al, one of the main areas in which alarms can be improved is alarm validation (ie, determining whether the alarm is actually valid) [4]. In this work, our main contribution is to this area. Our methodological approach to deal with alarm validation involves trying to fill the gap of having feasible solutions for mitigating the alarm fatigue problem by focusing on the issue of false positive alarms, which is known to be a serious problem that still remains unsolved.

Methods

Overview

With regard to methodology, we present a new approach to mitigate the alarm fatigue issue. We developed an application that attributes an FAP to alarms based on FAIs that we defined. Our reasoning algorithm uses the calculated FAP to decide whether to include an indication of FAP with a notification (ie, FAP_LABEL) before sending it to caregivers, in order to assist them in the complex task of choosing the next alarms to which they should respond.

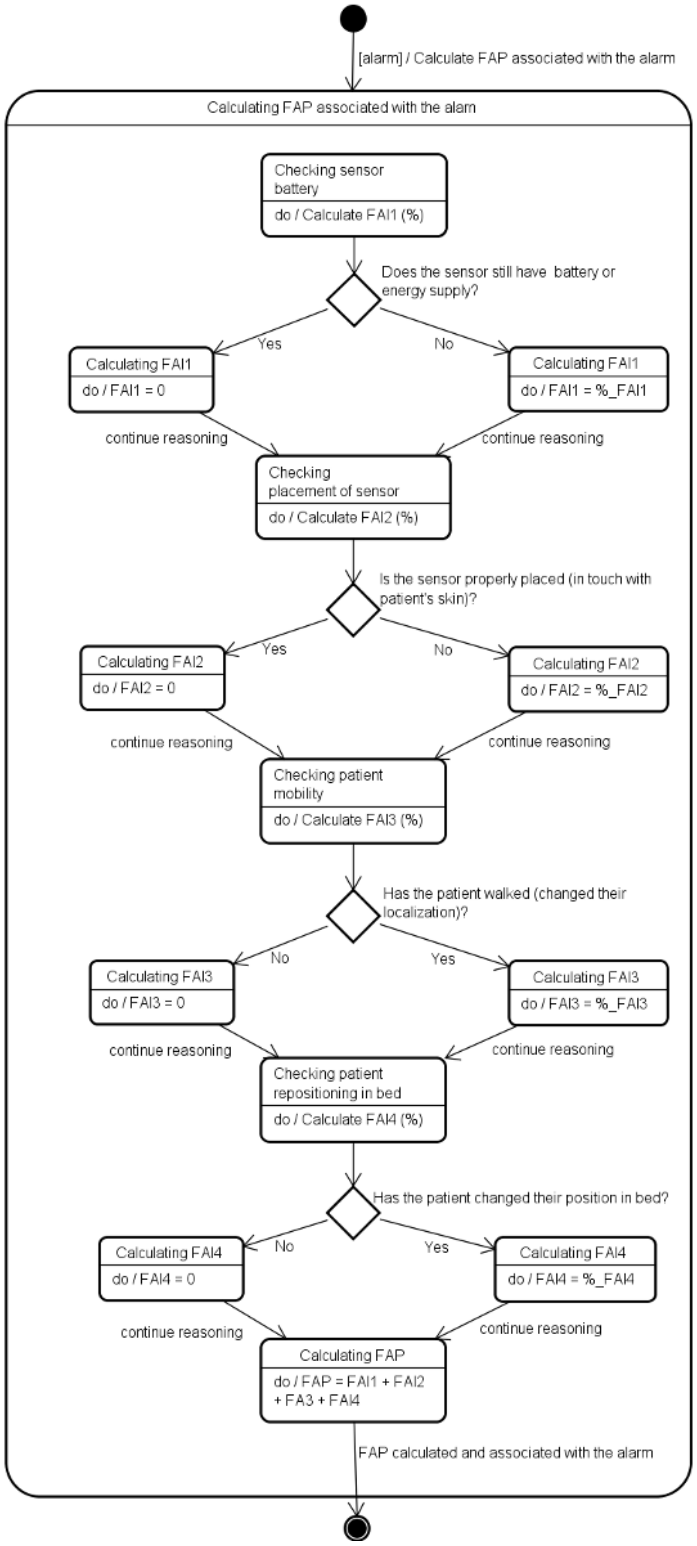
Reasoning Model for Deciding Whether to Include an FAP Label With a Notification

In our system, a notification is a type of message that is sent to caregivers and contains information about a detected alarm or a group of alarms. An FAP is associated with an individual alarm; we calculate the FAP according to the FAIs we describe next, while an FAP_LABEL, on the other hand, corresponds to the probability of a notification containing a false alarm.

We calculate the FAP of every alarm triggered by our system. However, the reasoning algorithm decides whether to include the indication of the FAP with a notification—as the FAP_LABEL—based on the FAIs. The FAP_LABEL is the piece of information that can be visualized by caregivers. The inputs for our algorithm are a notification and its context information, including information about the patient's conditions and sensors. After receiving these inputs, the reasoner starts working by analyzing the notification content and calculating the FAP_LABEL associated with it.

The processes to calculate the FAP and the FAP_LABEL are described below. Figure 1 presents a state machine diagram of the FAP reasoning process considering each alarm individually, as well as the reasoning modelling process that decides whether to notify caregivers through an FAP_LABEL indication.

Figure 1. State machine diagram showing how we calculate the false alarm probability (FAP) associated with an alarm. FAI: false alarm indicator.



Calculation of FAP Based on the FAIs

To calculate the FAP associated with each alarm, we defined four indicators of false alarms based on the information we gathered in our literature review. According to Kerr and Hayes, the main events that cause false alarms are patient movement or repositioning in bed and poor placement of sensors. Another common issue that triggers alarms is related to technical problems, such as lack of a battery in the monitoring devices [8].

The four FAIs defined in this case study represent information about (1) the duration of a sensor battery and the last time it was changed, (2) the last time the patient's skin was prepared to receive electrodes and the last time they were changed, (3) the patient's mobility, and (4) the patient's position in bed. To calculate the FAI percentage in our experiments, we considered

each indicator to have the same weight. The FAIs are listed below:

1. FAI1: Sensor battery FAI (SENSOR_BATTERY_FAI). This is an indication of the FAP associated with the battery-charge level of the sensors attached to the patient.
2. FAI2: Placement of sensor FAI (PLACEMENT_OF_SENSOR_FAI). FAI2 is related to the placement of a sensor (ie, whether a sensor is properly in touch with the patient's skin).
3. FAI3: Patient mobility FAI (PATIENT_MOBILITY_FAI). This indicator is related to patient mobility, which means that it can evaluate the probability that the alarm has been triggered due to his or her movement from the bed to other places.
4. FAI4: Patient repositioning FAI (PATIENT_REPOSITIONING_FAI). This indicator can be used to calculate the FAP related to patient repositioning (ie, whether the alarm has been sent simply because the patient may have changed his or her position in bed).

Inputs for Our Reasoning Algorithm Regarding Whether to Add an FAP_LABEL

As shown in Table 2, we defined eight inputs for our algorithm. There are four types of information that need to be manually inserted into our system by caregivers (Inputs 1-4), two types of data automatically collected via sensors (Inputs 5 and 7), and, finally, two inputs (Inputs 6 and 8) that are retrieved from the database by the system as historical patient data. Every input mentioned above is related to one of the four FAIs, as described below.

Table 2. Inputs for our reasoning algorithm.

Input	Input name	FAI[a] the input is used to calculate	Description	Type of related monitoring device
1	LEVEL_OF_ BATTERY	FAI1 (SENSOR_ BATTERY_FAI)	Level of battery for each monitoring device, including multiparametric monitors	Monitoring devices that use batteries
2	LAST_TIME_ BATTERY_ CHANGED	FAI1 (SENSOR_ BATTERY_FAI)	Last time the device's battery was changed	Monitoring devices that use batteries
3	LAST_TIME_ SKIN_ PREPARATION	FAI2 (PLACEMENT_OF_SENSOR_FAI)	Last time skin preparation occurred	Sensors that use electrodes
4	LAST_TIME_ ELEC-TRODES_ CHANGED	FAI2 (PLACEMENT_OF_SENSOR_FAI)	Last time electrodes were changed	Sensors that use electrodes
5	CURRENT_ PATIENT_ LO-CALIZATION	FAI3 (PATIENT_ MOBILITY_FAI)	The current patient's localization	Sensors used to track patient localization
6	LOG_LAST_ PATIENT_ LO-CALIZATION	FAI3 (PATIENT_ MOBILITY_FAI)	A log of the patient's last localization	Sensors used to track patient localization
7	CURRENT_ PATIENT_POSI-TION_IN_BED	FAI4 (PATIENT_REPOSITIONING_FAI)	The current position a patient occupies in a bed	Sensors used to track patient position in bed
8	LOG_LAST_ PATIENT_POSI-TIONS_IN_BED	FAI4 (PATIENT_REPOSITIONING_FAI)	The last positions a patient has occupied in a bed	Sensors used to track patient position in bed

[a]FAI: false alarm indicator.

Output of Our Reasoning Algorithm

There is one output of our algorithm—Output1: The probability that an alarm is false (ie, the FAP).

Application's Details: Technologies Utilized, Scenario, and Settings

To test our reasoning algorithm, we developed a system comprising an application (ie, the Producer App) that sends alarms to a broker who routes them to consumer applications that receive these alarms on behalf of the health care team. The system was developed in the Java language using the RabbitMQ message broker (Pivotal) [15]. The reason we decided to use RabbitMQ to handle the features related to data safety and scalability is to allow us to focus mainly on our functional requirements, since we are dealing with a high volume of alarms in our system.

Application Scenario

The application scenario consisted of a group of four patients being monitored in an ICU with sensors and monitoring devices, such as multiparametric monitors (see Figure 2), wearable devices, and external sensors that can be utilized with microcontrollers (see Figure 3).

Figure 2. The Philips Efficia CM100 monitor.

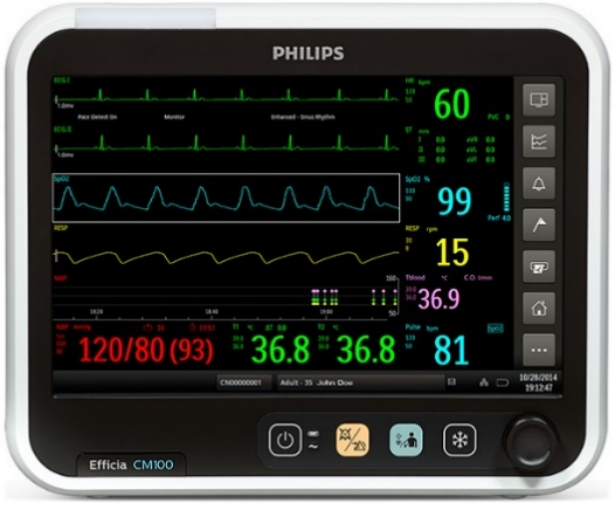


Figure 3. Arduino UNO microcontroller.

Monitoring Devices Used to Collect Biometric Patient Data

Philips Efficia CM100 Monitor

The Philips Efficia CM100 monitor [16] is commonly utilized to collect vital signs, such as electrocardiogram (ECG), breathing, temperature, noninvasive blood pressure (NIBP), oximetry (ie, peripheral oxygen saturation [SpO_2]), capnography (ie, end-tidal carbon dioxide [$EtCO_2$]), and invasive blood pressure (IBP).

eHealth Sensor Platform Kit

The electronic health (eHealth) Sensor Platform Complete Kit, version 2.0 (Cooking Hacks) [17] (see Figure 4), contains an eHealth Sensor Shield (Cooking Hacks; see Figure 5) compatible with the Arduino UNO (see Figure 3) [18] and Raspberry Pi (Raspberry Pi Foundation) [19] microcontrollers. It also contains 10 sensors to collect biometric data (see Figure 4): pulse, oxygen in blood, airflow (ie, breathing), body temperature, ECG, glucometer, galvanic skin response, blood pressure, patient position (ie, accelerometer), and muscle (ie, electromyography [EMG]).

Figure 4. eHealth Sensor Platform Complete Kit.

Figure 5. eHealth Sensor Shield.

Application Settings

In our simulated environment, patients were monitored through the use of two sensors: heart rate and temperature. The sensor readings were generated by a vital signs simulator that we developed. Regarding the sensor data simulated for each sensor, the temperature readings were generated randomly by the simulator within the 35.0°C-42.0°C range and the heart rate readings were randomly selected from the 40-188 beats-per-minute range. To define when a given temperature and heart rate reading represented an anomalous value that should trigger an alarm, we defined the thresholds shown in Table 3 for each patient.

Table 3. Defining the anomaly thresholds of temperature and heart rate sensors for each patient.

Patient ID	Minimum temperature, °C	Maximum temperature, °C	Minimum heart rate, BPM[a]	Maximum heart rate, BPM
1	35.5	39.0	60	100
2	35.0	38.5	55	95
3	35.5	39.5	60	100
4	35.5	38.5	50	100

[a]BPM: beats per minute.

In our experiments, we set the FAP_NOT_MIN at 75% (ie, the minimum value used as a reference to decide whether to add the FAP_LABEL to the notification). This means that every time the calculated FAP for an alarm was higher than or equal to 75%, our reasoner added the FAP_LABEL to the notification. Otherwise, we set the FAP_LABEL in our dataset to UNDEFINED, meaning that it was not included in the notification as an additional piece of information for caregivers (see Tables 4 and 5). We chose to use this strategy because we believe that only if this value is significant will it be useful to send the false alarm indication to the caregivers. Since we are working with an experimental version of our system, the choice of 75% for the FAP_NOT_MIN was selected arbitrarily. However, it is important to say that the medical staff can configure this value according to their preferences.

Results

In Tables 4 and 5 we present the results from our experiments. We illustrate a part of the output of our reasoning algorithm showing the first 10 notifications related to the temperature and heart rate vital signs, respectively. As one can see, FAP values were attributed to the alarms, and the FAP_LABELs were added to notifications by the reasoner. The first four columns represent the Notification ID (NID), Ward ID (WID), Patient ID (PID), and Alarm ID (AID), respectively.

Table 4. Results of our experiments for notifications related to temperature alarms.

NID[a]	WID[b]	PID[c]	AID[d]	Sensor value, °C	Alarm timestamp, date and time	FAP[e],%	Notification timestamp, date and time	FAP_LABEL, %
1	1	1	1	35.0	2019-07-02 21:51:06.291	50.0	2019-07-02 21:51:06.334	UNDEFINED
2	1	4	2	42.0	2019-07-02 21:51:08.328	25.0	2019-07-02 21:51:08.328	UNDEFINED
3	1	3	4	41.0	2019-07-02 21:51:12.457	50.0	2019-07-02 21:51:12.457	UNDEFINED
4	1	2	9	41.0	2019-07-02 21:51:43.223	75.0	2019-07-02 21:51:43.223	75.0
5	1	1	12	42.0	2019-07-02 21:52:03.697	50.0	2019-07-02 21:56:06.334	UNDEFINED
5	1	1	15	42.0	2019-07-02 21:52:20.053	100.0	2019-07-02 21:56:06.334	100.0
5	1	1	16	41.0	2019-07-02 21:52:24.135	75.0	2019-07-02 21:56:06.334	75.0
5	1	1	17	35.0	2019-07-02 21:52:32.309	25.0	2019-07-02 21:56:06.334	UNDEFINED
5	1	1	18	42.0	2019-07-02 21:52:42.594	50.0	2019-07-02 21:56:06.334	UNDEFINED
5	1	1	20	41.0	2019-07-02 21:52:50.774	50.0	2019-07-02 21:56:06.334	UNDEFINED

[a]NID: Notification ID.

[b]WID: Ward ID.

[c]PID: Patient ID.

[d]AID: Alarm ID.

[e]FAP: false alarm probability.

Table 5. Results of our experiments for notifications related to heart rate vital signs.

NID[a]	WID[b]	PID[c]	AID[d]	Sensor value, BPM[e]	Alarm timestamp, date and time	FAP[f],%	Notification timestamp, date and time	FAP_LABEL, %
1	1	2	1	108.0	2019-07-02 21:51:09.375	75.0	2019-07-02 21:51:09.39	75.0
2	1	1	2	145.0	2019-07-02 21:51:11.432	25.0	2019-07-02 21:51:11.432	UNDEFINED
3	1	4	6	123.0	2019-07-02 21:51:21.721	50.0	2019-07-02 21:51:21.722	UNDEFINED
4	1	3	8	116.0	2019-07-02 21:51:25.827	50.0	2019-07-02 21:51:25.827	UNDEFINED
5	1	2	3	156.0	2019-07-02 21:51:15.539	0.0	2019-07-02 21:56:09.397	UNDEFINED
5	1	2	5	159.0	2019-07-02 21:51:19.667	50.0	2019-07-02 21:56:09.397	UNDEFINED
5	1	2	7	44.0	2019-07-02 21:51:23.776	75.0	2019-07-02 21:56:09.397	75.0
5	1	2	9	164.0	2019-07-02 21:51:27.874	50.0	2019-07-02 21:56:09.397	UNDEFINED
5	1	2	16	184.0	2019-07-02 21:51:44.254	25.0	2019-07-02 21:56:09.397	UNDEFINED
5	1	2	23	51.0	2019-07-02 21:52:00.641	0.0	2019-07-02 21:56:09.397	UNDEFINED

[a]NID: Notification ID.

[b]WID: Ward ID.

[c]PID: Patient ID.

[d]AID: Alarm ID.

[e]BPM: beats per minute.

[f]FAP: false alarm probability.

Discussion

Alarm safety is a complex problem to solve, influenced by a number of factors that extrapolate technology challenges and limitations, such as human influences, difficult patient conditions, a wide variety of environmental conditions, and even staffing cultures [12]. Alarm hazards are still a big challenge for members of the health care teams in ICUs. As practice settings continue to become more technology driven, effective interventions for alarm hazards in ICU settings are crucial. Feasible strategies should be provided in order to allow nurses to respond to the call to ensure patient safety in an increasingly complex care environment [10].

In this work, we tried to fill the gap of having feasible solutions to mitigate the alarm fatigue problem by focusing on the issue of false positive alarms, known to be a serious problem that yet remains unsolved. This paper presented a reasoning algorithm to detect false alarms based on alarm-context information provided automatically by the use of sensors and wearable devices and manually by the inputs of caregivers.

In our experiments, we created a database of simulated alarm-context information to establish a basis for the development of our algorithm in order to confirm H1 and H2 in experimental settings. As we can see in the FAP column of Tables 4 and 5, every alarm generated by the sensors and monitoring devices in our experiments had an FAP value associated with it by our reasoning algorithm. Our algorithm also added an indication of an FAP (ie, FAP_LABEL) to the notifications sent to caregivers. This information is available in the FAP_LABEL column of our dataset (see Tables 4 and 5).

Regarding H3, which declares that patient safety will not be compromised if and when the reasoning algorithm decides to add an FAP_LABEL to the notification, we can assume that is confirmed, since our algorithm does not stop an alarm from being triggered even when the calculated FAP is considered very high. We can see an example of this information in the sixth row of Table 4, where the alarm (ie, AID=15) still triggered a notification (ie, NID=5), even though it had a calculated FAP of 100%.

As future work, we are planning to evolve our solution to support an optimized version of our reasoning algorithm that calculates the optimal FAP_NOT_MIN based on the real-time volume of alarms being triggered in an ICU.

Another plan for future work is to develop a machine learning–based algorithm capable of predicting both the FAP and FAP_LABEL based on a dataset that contains the ICU information history, such as patients' conditions, sensors, and alarms.

Acknowledgments

This work was supported by grants from two Brazilian government agencies: CNPq (National Council of Research Development) and FAPERJ (Foundation for Research Support of the State of Rio de Janeiro).

Conflicts of Interest

None declared.

References

1. Fernandes CO, Pereira De Lucena CJ. A software framework for remote patient monitoring by using multi-agent systems support. JMIR Med Inform 2017 Mar 27;5(1):e9 [FREE Full text] [doi: 10.2196/medinform.6693] [Medline: 28347973]
2. Fernandes CO, Pereira De Lucena CJ, de Souza e Silva D. Smart depth of anesthesia monitoring with EEG sensors and agent-based technology. In: Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). New York, NY: IEEE; 2017 Aug 04 Presented at: 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI); August 4-8, 2017; San Francisco, CA p. 1-8 URL: https://doi.org/10.1109/UIC-ATC.2017.8397455 [doi: 10.1109/uic-atc.2017.8397455]
3. Fernandes CO, Pereira de Lucena CJ, Pereira de Lucena CA, Alvares de Azevedo B. Enabling a smart and distributed communication infrastructure in healthcare. In: Chen YW, Torro C, Tanaka S, Howlett RC, Jain L, editors. Innovation in Medicine and Healthcare 2015. Smart Innovation, Systems and Technologies, vol 45. Cham, Switzerland: Springer International Publishing; Aug 12, 2016:435-446.
4. Imhoff M, Kuhls S, Gather U, Fried R. Smart alarms from medical devices in the OR and ICU. Best Pract Res Clin Anaesthesiol 2009 Mar;23(1):39-50. [doi: 10.1016/j.bpa.2008.07.008] [Medline: 19449615]
5. Jones K. Alarm fatigue a top patient safety hazard. CMAJ 2014 Feb 18;186(3):178 [FREE Full text] [doi: 10.1503/cmaj.109-4696] [Medline: 24418978]
6. Lawless ST. Crying wolf. Crit Care Med 1994;22(6):981-985 [FREE Full text] [doi: 10.1097/00003246-199406000-00017]
7. Sendelbach S. Alarm fatigue. Nurs Clin North Am 2012 Sep;47(3):375-382. [doi: 10.1016/j.cnur.2012.05.009] [Medline: 22920428]
8. Kerr JH, Hayes B. An "alarming" situation in the intensive therapy unit. Intensive Care Med 1983 May;9(3):103-104. [doi: 10.1007/bf01772574]
9. Cvach M. Monitor alarm fatigue: An integrative review. Biomed Instrum Technol 2012;46(4):268-277. [doi: 10.2345/0899-8205-46.4.268] [Medline: 22839984]
10. Tanner T. The problem of alarm fatigue. Nurs Womens Health 2013;17(2):153-157. [doi: 10.1111/1751-486X.12025] [Medline: 23594329]
11. ECRI. URL: https://www.ecri.org/ [accessed 2020-05-07]
12. Keller JP. Clinical alarm hazards: A "top ten" health technology safety concern. J Electrocardiol 2012;45(6):588-591. [doi: 10.1016/j.jelectrocard.2012.08.050] [Medline: 23022300]
13. Winters BD, Cvach MM, Bonafide CP, Hu X, Konkani A, O'Connor MF, Society for Critical Care Medicine Alarm and Alert Fatigue Task Force. Technological distractions (Part 2): A summary of approaches to manage clinical alarms with intent to reduce alarm fatigue. Crit Care Med 2018 Jan;46(1):130-137. [doi: 10.1097/CCM.0000000000002803] [Medline: 29112077]
14. Fernandes CO, Miles S, Pereira De Lucena CJ, Cowan D. Artificial intelligence technologies for coping with alarm fatigue in hospital environments because of sensory overload: Algorithm development and validation. J Med Internet Res 2019 Nov 26;21(11):e15406 [FREE Full text] [doi: 10.2196/15406] [Medline: 31769762]
15. RabbitMQ. URL: https://www.rabbitmq.com/ [accessed 2020-05-07] [WebCite Cache ID 78odvPSaX]
16. Koninklijke Philips. Soluções de monitoração Efficia e cuidados com o paciente (Efficia monitoring solutions and patient care) [webpage in Portuguese] URL: https://www.philips.com.br/healthcare/solutions/value-products/efficia [accessed 2020-05-07]
17. Cooking Hacks. e-Health Sensor Platform V2.0 for Arduino and Raspberry Pi [biometric / medical applications] URL: https://www.cooking-hacks.com/documentation/tutorials/ehealth-biometric-sensor-platform-arduino-raspberry-pi-medical [accessed 2020-05-07] [WebCite Cache ID 6jxCIvFSO]
18. Arduino. URL: https://www.arduino.cc/ [accessed 2020-05-07] [WebCite Cache ID 6jwDCGjRi]

19. Raspberry Pi. URL: https://www.raspberrypi.org/ [accessed 2020-05-07]

Abbreviations

AI: artificial intelligence
AID: Alarm ID
CNPq: National Council of Research Development
ECG: electrocardiogram
ECRI: Emergency Care Research Institute
eHealth: electronic health
EMG: electromyography
EtCO2: end-tidal carbon dioxide
FAI: false alarm indicator
FAP: false alarm probability
FAPERJ: Foundation for Research Support of the State of Rio de Janeiro
H1: Hypothesis 1
H2: Hypothesis 2
H3: Hypothesis 3
IBP: invasive blood pressure
ICU: intensive care unit
NIBP: noninvasive blood pressure
NID: Notification ID
OR: operating room
PID: Patient ID
SpO2: peripheral oxygen saturation
WID: Ward ID

Edited by G Eysenbach; submitted 08.07.19; peer-reviewed by M Cvach, L Roa, H Yu, S Monteiro, D Koutsouris, I Pires; comments to author 28.08.19; revised version received 15.10.19; accepted 26.01.20; published 20.05.20.

Please cite as:
Fernandes C, Miles S, Lucena CJP
Detecting False Alarms by Analyzing Alarm-Context Information: Algorithm Development and Validation
JMIR Med Inform 2020;8(5):e15407
URL: http://medinform.jmir.org/2020/5/e15407/
doi:10.2196/15407
PMID:32432551

Original Paper

Clinical Desire for an Artificial Intelligence–Based Surgical Assistant System: Electronic Survey–Based Study

Soo Jin Park[1], MD; Eun Ji Lee[1], MD; Se Ik Kim[1], MD; Seong-Ho Kong[2], MD, PhD; Chang Wook Jeong[3], MD, PhD; Hee Seung Kim[1], MD, PhD

[1]Department of Obstetrics and Gynecology, Seoul National University College of Medicine, Seoul, Republic of Korea

[2]Department of Surgery, Seoul National University College of Medicine, Seoul, Republic of Korea

[3]Department of Urology, Seoul National University College of Medicine, Seoul, Republic of Korea

Corresponding Author:
Hee Seung Kim, MD, PhD
Department of Obstetrics and Gynecology
Seoul National University College of Medicine
101 Daehak-Ro
Jongno-Gu
Seoul, 03080
Republic of Korea
Phone: 82 02 2072 4863
Email: bboddi0311@gmail.com

Abstract

Background: Techniques utilizing artificial intelligence (AI) are rapidly growing in medical research and development, especially in the operating room. However, the application of AI in the operating room has been limited to small tasks or software, such as clinical decision systems. It still largely depends on human resources and technology involving the surgeons' hands. Therefore, we conceptualized AI-based solo surgery (AISS) defined as laparoscopic surgery conducted by only one surgeon with support from an AI-based surgical assistant system, and we performed an electronic survey on the clinical desire for such a system.

Objective: This study aimed to evaluate the experiences of surgeons who have performed laparoscopic surgery, the limitations of conventional laparoscopic surgical systems, and the desire for an AI-based surgical assistant system for AISS.

Methods: We performed an online survey for gynecologists, urologists, and general surgeons from June to August 2017. The questionnaire consisted of six items about experience, two about limitations, and five about the clinical desire for an AI-based surgical assistant system for AISS.

Results: A total of 508 surgeons who have performed laparoscopic surgery responded to the survey. Most of the surgeons needed two or more assistants during laparoscopic surgery, and the rate was higher among gynecologists (251/278, 90.3%) than among general surgeons (123/173, 71.1%) and urologists (35/57, 61.4%). The majority of responders answered that the skillfulness of surgical assistants was "very important" or "important." The most uncomfortable aspect of laparoscopic surgery was unskilled movement of the camera (431/508, 84.8%) and instruments (303/508, 59.6%). About 40% (199/508, 39.1%) of responders answered that the AI-based surgical assistant system could substitute 41%-60% of the current workforce, and 83.3% (423/508) showed willingness to buy the system. Furthermore, the most reasonable price was US $30,000-50,000.

Conclusions: Surgeons who perform laparoscopic surgery may feel discomfort with the conventional laparoscopic surgical system in terms of assistant skillfulness, and they may think that the skillfulness of surgical assistants is essential. They desire to alleviate present inconveniences with the conventional laparoscopic surgical system and to perform a safe and comfortable operation by using an AI-based surgical assistant system for AISS.

(JMIR Med Inform 2020;8(5):e17647) doi:10.2196/17647

KEYWORDS

artificial intelligence; solo surgery; laparoscopic surgery

Introduction

Artificial intelligence (AI) has been rapidly developing in recent years, and relevant research is being actively conducted in the health care field through deep learning and big data technology [1]. AI applied in the medical area can be divided into the following two categories: virtual and physical AI. Virtual AI includes the programs that can help clinical diagnosis, whereas physical AI involves smart operating rooms, nanorobots, and patient-assistance systems [2]. In particular, physical AI in the operating room can assist the operator or replace the assistant during surgery [2,3]. For instance, the da Vinci surgical system, which is the first computer-based robotic surgical system approved by the US Food and Drug Administration in 2000, has been widely used for minimally invasive surgery, including laparoscopic surgery. The demand for the robotic surgical system is rapidly increasing in the surgical areas of gynecology, general surgery, and urology [4]. This increase in demand is due to reduced surgeon fatigue and improved surgical access through ergonomic instruments and three-dimensional imaging [4,5].

However, the current robotic surgical system still depends on coordination of the human eye and hand, which is insufficient in terms of autonomy or interaction [6,7]. In particular, the injection of carbon dioxide and insertion of trocars into the peritoneal cavity are still performed by surgeons without the aid of a robotic surgical system, and the laparoscopic camera and instruments are adjusted manually to the target by surgeons. Thus, an automated robotic surgical system that is better than the current master-slave approach may be expected to reduce human error and thereby improve the quality of surgery. Up to now, relevant studies have mainly focused on the development of robots capable of performing short surgical tasks, such as knot tie and needle insertion [8,9], and the application of voice interaction technology during surgery may be one of the crucial elements that should be developed in an AI-based surgical assistant system [10-12].

Nevertheless, high medical cost may be one of the barriers to the adoption of an AI-based surgical assistant system [13], and it is not yet know how this system will improve the quality of surgery or reduce human resources effectively. Therefore, we conceptualized AI-based solo surgery (AISS) that was defined as laparoscopic surgery conducted by only one surgeon with support from an AI-based surgical assistant system and considered the clinical desire for AISS via an electronic survey (e-survey).

An e-survey has been a common method of research in human and social sciences since the 1990s. In the case of research using a web-based questionnaire, it is possible to attach pictures or materials in order to avoid response omission as much as possible and avoid inconsistent or out-of-frame results. Besides, data can be effectively organized and archived without paper resources, and distribution via email can be quickly done through a URL [14]. Moreover, by distributing the web questionnaire via email, it is possible to limit the target respondents to people belonging to a specific community so that the questionnaire survey is conducted for experts in the relevant field.

Therefore, we performed an e-survey to investigate the clinical desire for an AI-based surgical assistant system for AISS as compared with the current laparoscopic surgical system and to determine the reasonable cost of such an AI-based surgical assistant system for AISS.

Methods

Survey

We surveyed gynecologists from the Korean Society of Obstetrics and Gynecology, urologists from the Korean Urologic Association, and general surgeons from the Korean Surgical Society between June and August 2017 through nownsurvey (ELIMNET Co, Ltd) [15], a commercially available e-survey platform. In this survey, the AI-based surgical assistant system for AISS was considered to have the following functions: camera automatic recognition and operation function through voice commands; action as an assistant by manipulating surgical instruments through automatic screen recognition and voice commands; and smart storage for recognizing, indexing, and storing surgical procedures while recording specific events. There were a total of 13 questions that included six items about the responder's experience, two about limitations of the conventional laparoscopic surgical system, and five about the clinical desire for an AI-based surgical assistant system for AISS (Table 1). We estimated that 5000 gynecologists, 7000 general surgeons, and 2500 urologists would participate in the survey. This study was approved by the Institutional Review Board of Seoul National University Hospital (approval no: 1910-131-1072).

Table 1. Questionnaire details.

Variable and question number	Question
Experience	
1	What type of hospital do you work at?
2	What department do you work in?
3	How many patients do you perform laparoscopic surgery in monthly?
4	How many assistants do you need during laparoscopic surgery?
5	What kinds of assistants do you want during laparoscopic surgery?
6	How important is the skillfulness of your assistant for successful laparoscopic surgery?
Limitation	
7	What are your discomforts during laparoscopic surgery owing to inexperienced camera assistants? (multiple choice)
8	What are your discomforts during laparoscopic surgery owing to inexperienced laparoscopic instrument assistants? (multiple choice)
Desire	
9	What functions do you expect to be included in the AI[a]-based surgical assistant system for AISS[b]? (multiple choice)
10	What percentage of your assistant's function will the AI-based surgical assistant system for AISS replace?
11	Would you want to buy the AI-based surgical assistant system for AISS if it thrives?
12	Why would you want to buy the AI-based surgical assistant system for AISS? (multiple choice)
13	How much would you like to pay for the AI-based surgical assistant system for AISS?

[a]AI: artificial intelligence.

[b]AISS: artificial intelligence–based solo surgery.

Data Analysis

We analyzed each question by using descriptive statistics. Additionally, we analyzed all the respondents, and the response rate was 3.5%. Each item in the questionnaire was stratified according to the surgeons' fields as follows: gynecologists, urologists, and general surgeons. Categorical variables were analyzed with the chi-square test or Fisher exact test using the statistical software SPSS 20.0 (IBM Corp, Armonk, New York, USA). A *P* value <.05 was considered statistically significant.

Results

Experience

Table 2 shows the demographic data of the responders. A total of 508 people responded to the questionnaire, and there were 278 gynecologists, 173 general surgeons, and 57 urologists. Among the three surgeon fields, most of the urologists (49/57, 86.0%) worked at a university hospital, whereas relatively many gynecologists (67/278, 24.1%) worked as general practitioners. Moreover, most of the urologists (43/57, 75.4%) performed laparoscopic surgery in less than 10 cases per month, whereas relatively many general surgeons performed laparoscopic surgery in 31 or more cases per month (40/173, 23.1%). In terms of the number of assistants during laparoscopic surgery, 38.6% (22/57) of urologists required one or less assistant, whereas 90.3% (251/278) of gynecologists required two or more assistants.

In terms of the preferred assistant during laparoscopic surgery, most of the urologists (33/57, 57.9%) preferred fellows, whereas many general surgeons (76/173, 43.9%) preferred physician assistants (Figure 1). With regard to the importance of the skillfulness of assistants, who manipulate cameras or instruments, for successful laparoscopic surgery, most of the responders indicated "very important" or "important," regardless of the surgeon field. Although the trend was similar among the three surgeon fields with regard to the camera assistant, general surgeons (33/173, 19.1%) relatively underestimated the importance of the skillfulness of instrument assistants as compared with gynecologists (93/278, 33.5%) or urologists (18/57, 31.6%) (Figure 2).

Table 2. Demographic data.

Answers	Total (N=508), n (%)	Gynecologists (N=278), n (%)	General surgeons (N=173), n (%)	Urologists (N=57), n (%)	*P* value
Working hospital					**<.001**
University hospital	306 (60.2)	146 (52.5)	111 (64.2)	49 (86.0)	
General hospital	76 (15.0)	34 (12.2)	36 (20.8)	6 (10.5)	
Semi hospital	49 (9.6)	31 (11.2)	17 (9.8)	1 (1.8)	
General practitioner	77 (15.2)	67 (24.1)	9 (5.2)	1 (1.8)	
Total number of laparoscopic surgeries per month					**<.001**
0-10	232 (45.7)	136 (48.9)	53 (30.6)	43 (75.4)	
11-30	181 (35.6)	89 (32.0)	80 (46.2)	12 (21.1)	
≥31	95 (18.7)	53 (19.1)	40 (23.1)	2 (3.5)	
Number of assistants during laparoscopic surgery					**<.001**
0	1 (0.2)	0 (0.0)	1 (0.6)	0 (0.0)	
1	98 (19.3)	27 (9.7)	49 (28.3)	22 (38.6)	
2	349 (68.7)	214 (77.0)	106 (61.3)	29 (50.9)	
≥3	60 (11.8)	37 (13.3)	17 (9.8)	6 (10.5)	

Figure 1. Comparison of assistants preferred during laparoscopic surgery.

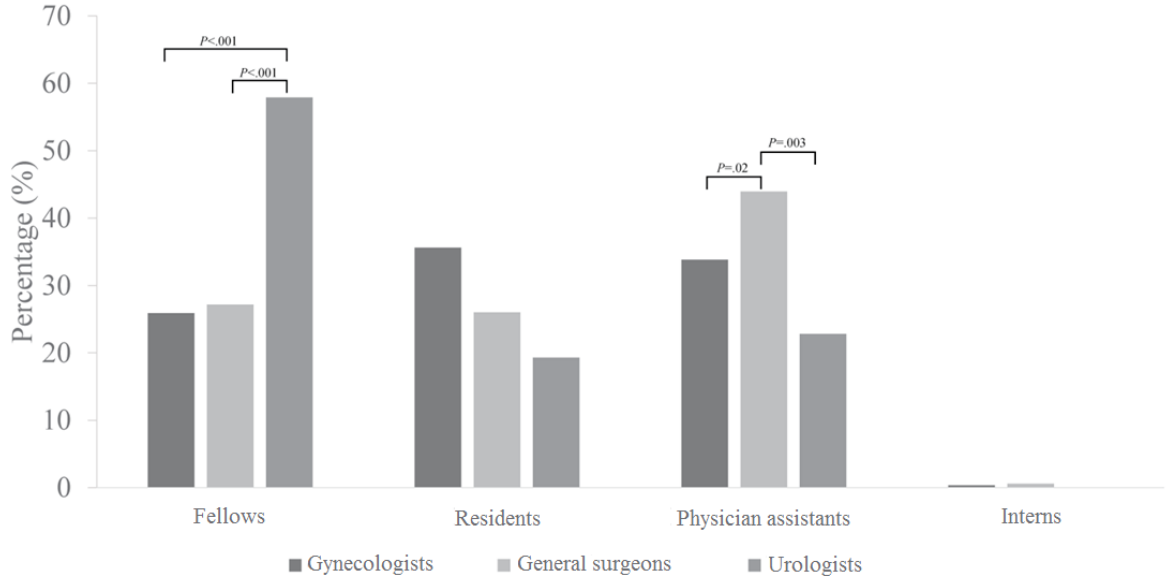
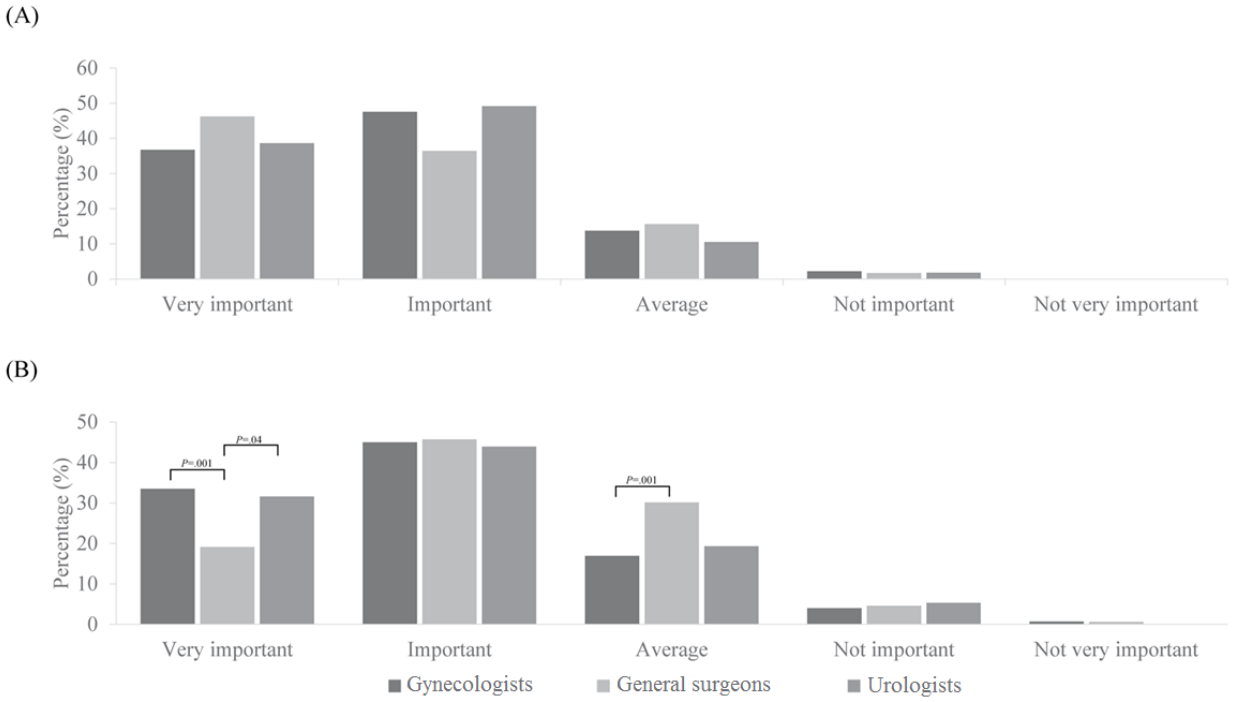


Figure 2. Comparison of the importance of the skillfulness of (A) camera and (B) instrument assistants.



Limitation

Table 3 shows the responses to questions on the surgeons' discomforts related to inexperienced camera and instrument assistants for the conventional laparoscopic surgical system. With regard to the camera assistant, 84.8% (431/508) of the responders were unsatisfied with unskilled movement of the camera in the intended direction. In particular, gynecologists (69/278, 24.8%) had more complaints about contamination of the camera lens by blood or body fluid as compared with general surgeons (26/173, 15.0%) or urologists (6/57, 10.5%). With regard to the instrument assistant, 59.6% (303/508) of the responders were unsatisfied with unskilled movement of the instruments in the intended direction. In particular, general surgeons (103/173, 59.5%) had more complaints about tissue damage or bleeding by inappropriate traction and urologists (24/57, 42.1%) had more complaints about collision between the instruments as compared with the other surgeons. On the other hand, gynecologists (32/278, 11.5%) had more complaints about swaying of the instruments as compared with the other surgeons.

Table 3. Surgeons' discomforts regarding the conventional laparoscopic surgical system.

Discomfort	Total (N=508), n (%)	Gynecologists (N=278), n (%)	General surgeons (N=173), n (%)	Urologists (N=57), n (%)	*P* value
Camera assistant					
Unskilled movement of the camera in the intended direction	431 (84.8)	227 (81.7)	151 (87.3)	53 (93.0)	.05
Dizziness due to excessive camera movement	174 (34.3)	96 (34.5)	61 (35.3)	17 (29.8)	.75
Inappropriate field of view due to excessive zoom in or out	156 (30.7)	80 (28.8)	59 (34.1)	17 (29.8)	.49
Condensation on the camera lens	123 (24.2)	76 (27.3)	39 (22.5)	8 (14.0)	.08
Contamination of the camera lens by blood or body fluid	101 (19.9)	69 (24.8)	26 (15.0)	6 (10.5)	.01
Blurriness of the camera	101 (19.9)	57 (20.5)	32 (18.5)	12 (21.1)	.85
Instrument assistant					
Unskilled movement of the instruments in the intended direction	303 (59.6)	172 (61.9)	94 (54.3)	37 (64.9)	0.20
Inappropriate tissue traction due to lack of power to pull or push	196 (38.6)	102 (36.7)	79 (45.7)	15 (26.3)	0.02
Dangerous movement of the instruments outside the camera view	145 (28.5)	75 (27.0)	47 (27.2)	23 (40.4)	0.11
Tissue damage or bleeding by inappropriate traction	189 (37.2)	63 (22.7)	103 (59.5)	23 (40.4)	<.001
Collision between the instruments	126 (24.8)	57 (20.5)	45 (26.0)	24 (42.1)	.002
Insufficient removal of intra-abdominal smoke during surgery	77 (15.2)	52 (18.7)	18 (10.4)	7 (12.3)	.047
Swaying of the instruments	44 (8.7)	32 (11.5)	9 (5.2)	3 (5.3)	.04

Desire

Table 4 depicts the functions that should be included in an AI-based surgical assistant system for AISS to overcome the limitations of the current laparoscopic surgical system. More than half of the responders preferred intuitive and easy maneuverability (308/508, 60.6%), a demister and self-cleaning system for the laparoscopic camera lens (326/508, 64.2%), and safety for minimizing tissue damage (279/508, 54.9%). In particular, more urologists (29/57, 50.9%) desired fast running by minimizing time delay as compared with gynecologists (86/278, 30.9%) and general surgeons (67/173, 38.7%). However, interest in the autosave or voice command system for special events during the operation was the lowest among the three surgeon fields. In terms of the possibility that the AI-based surgical assistant system for AISS can replace the functions of assistants, about 40% (199/508, 39.1%) of responders expected it to substitute 41%-60% of the existing workforce (Figure 3).

When asked about the purchase intention and reasonable price to buy the AI-based surgical assistant system for AISS, 83.3% (423/508) of all responders wanted to buy the system. The most common reason for wanting to buy the system was the comfort of laparoscopic surgery (257/508, 50.6%). In particular, general surgeons had a relatively strong desire to decrease the burden of repetitive training for assistants, whereas they had less interest in the reduction of the operation time by purchasing the AI-based surgical assistant system for AISS as compared with gynecologists. Regarding the reasonable price for the system, 29.7% (151/508) of the responders had a willingness to pay US $30,000-50,000 (Table 5).

Table 4. Functions that should be included in an artificial intelligence–based surgical assistant system.

Function	Total (N=508), n (%)	Gynecologists (N=278), n (%)	General surgeons (N=173), n (%)	Urologists (N=57), n (%)	*P* value
Intuitive and easy maneuverability	308 (60.6)	164 (59.0)	108 (62.4)	36 (63.2)	.71
Demister and self-cleaning system for the laparoscopic camera lens	326 (64.2)	186 (66.9)	108 (62.4)	32 (56.1)	.26
Safety for minimizing tissue damage	279 (54.9)	150 (54.0)	97 (56.1)	32 (56.1)	.89
Reasonable size of the instruments avoiding operator disturbance	248 (48.8)	130 (46.8)	86 (49.7)	32 (56.1)	.42
Stabilization of the laparoscopic camera and instruments	220 (43.3)	122 (43.9)	69 (39.9)	29 (50.9)	.33
Stable movements not causing dizziness	208 (40.9)	114 (41.0)	65 (37.6)	29 (50.9)	.21
Functions for complex movements, such as axial rotation of the 30-degree camera and manipulation of the flexible scope	213 (41.9)	107 (38.5)	80 (46.2)	26 (45.6)	.22
Fast running by minimizing time delay	182 (35.8)	86 (30.9)	67 (38.7)	29 (50.9)	.01
Autosave or voice command system for special events during the operation	139 (27.4)	76 (27.3)	46 (26.6)	17 (29.8)	.89

Figure 3. Expectations about how much an artificial intelligence–based surgical assistant system can replace the existing workforce.

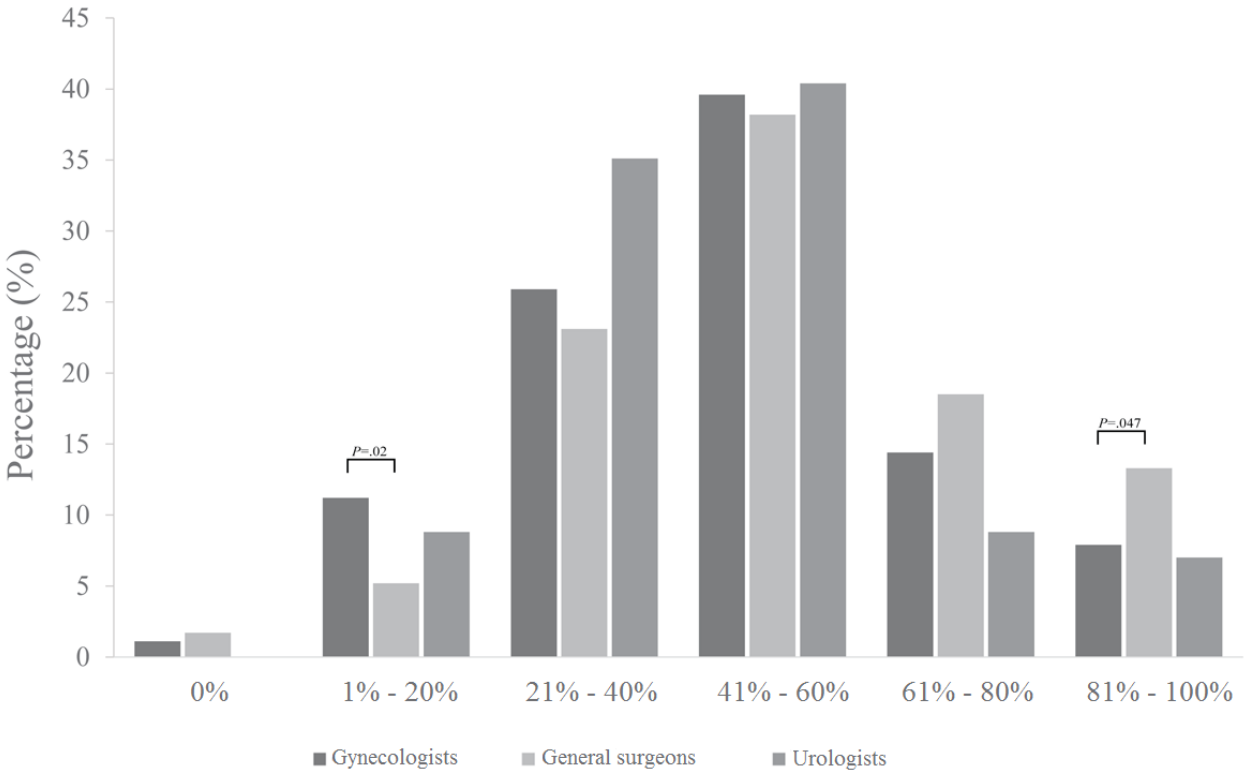


Table 5. Purchase intention and reasonable price to buy the artificial intelligence–based surgical assistant system.

Answers	Total (N=508), n (%)	Gynecologists (N=278), n (%)	General surgeons (N=173), n (%)	Urologists (N=57), n (%)	*P* value
Purchase intention	423 (83.3)	222 (79.9)	151 (87.3)	50 (87.7)	.08
Reason to buy the system					
Comfort of laparoscopic surgery	257 (50.6)	142 (64.0)	84 (55.6)	31 (62.0)	.27
Improved safety and maturity of laparoscopic surgery	245 (48.2)	126 (56.8)	85 (56.3)	34 (68.0)	.31
Decreased number of assistants	204 (40.2)	101 (45.5)	75 (49.7)	28 (56.0)	.37
Decreased burden of repetitive training for assistants	197 (38.8)	95 (42.8)	82 (54.3)	20 (4.0)	.06
Reduced operation time	119 (23.4)	74 (33.3)	27 (17.9)	18 (36.0)	.002
Improved convenience of research based on the autosave function	114 (22.4)	59 (26.6)	39 (25.8)	16 (32.0)	.68
Reasonable price (US$)					**.04**
<30,000	87 (17.1)	61 (21.9)	21 (12.1)	5 (8.8)	
30,000-50,000	151 (29.7)	83 (29.9)	53 (30.6)	15 (26.3)	
50,000-100,000	139 (27.4)	77 (27.7)	43 (24.9)	19 (33.3)	
100,000-150,000	79 (15.6)	33 (11.9)	36 (20.8)	10 (17.5)	
150,000-200,000	28 (5.5)	11 (4.0)	13 (7.5)	4 (7.0)	
≥200,000	24 (4.7)	13 (4.7)	7 (4.0)	4 (7.0)	

Discussion

Principal Findings

This study involved a survey about the clinical desire for an AI-based surgical assistant system for AISS among surgeons who currently perform laparoscopic surgery. In this survey, we identified the importance of assistants and the discomforts with the conventional laparoscopic surgical system and determined surgeons' expectations and demands for new AI-based robotic surgery aids.

Experience

In terms of experience, gynecologists were more likely to have two assistants than general surgeons and urologists. The reason is that gynecologists may use a uterine manipulator frequently during laparoscopic gynecologic surgery [16]. Therefore, gynecologists commonly require two or more assistants for laparoscopic surgery, including two assistants who hold a laparoscopic camera and a uterine manipulator.

On the other hand, urologists' preference for fellows as surgical assistants could be related to more common practice in university hospitals. Moreover, urologists can be less dependent on residents during surgery, which may be similar for general surgeons who prefer physician assistants as surgical assistants. Furthermore, most of the responders valued the skillfulness of surgical assistants who manipulate the laparoscopic camera and instrument assistants, because the extent of assistant experience may be closely related to the operation time and complication rate [17]. Recently, in the Republic of Korea, owing to the implementation of the special act regarding an 80-hour workweek for residents, their working time has reduced, and thereby, the number of cases of surgical training has reduced [18]. In contrast, physician assistants are still useful for coordination in the operating room because of their high level of proficiency based on repetitive work [19]. Therefore, most surgeons seem to prefer fellows or physician assistants who are proficient in laparoscopic surgery rather than residents or interns because of their skillfulness as surgical assistants in the Republic of Korea.

Limitation

In terms of limitation, most of the surgeons felt uncomfortable with camera assistants when they showed unskilled movement of the camera in the intended direction and instrument assistants when they showed unskilled movement of the instruments in the intended direction. This result is consistent with the finding that most of the surgeons considered the skillfulness of surgical assistants as "very important" or "important," regardless of the field.

Desire

In terms of desire, the essential functions desired to be present in an AI-based surgical assistant system for AISS were intuitive and easy maneuverability, a demister and self-cleaning system for the laparoscopic camera lens, and safety for minimizing tissue damage. Interestingly, only 10%-20% of surgeons complained about discomfort regarding the camera lens or foreign objects, whereas a high percentage of surgeons desired a self-cleaning system for AISS. These findings seem to be associated with the role of surgical assistants in camera cleaning when using the conventional laparoscopic surgical system, which is perceived as an essential function by the operator, and the absence of an uncomfortable feeling with the current system.

Notably, more than 80% of the responders intended to buy the AI-based surgical assistant system for AISS, and the reasons for buying it were comfort of laparoscopic surgery and improved safety and maturity of laparoscopic surgery. Considering the results from the questions on the conventional laparoscopic surgical system, surgeons showed a tendency to overcome current constraints regarding laparoscopic surgery with the AI-based surgical assistant system, especially with regard to the skillfulness of assistants.

The majority of responders anticipated that the introduction of the AI-based surgical assistant system would replace the existing workforce by 41%-60%. Therefore, an AI-based surgical assistant system for AISS could be a great solution in university hospitals where resident working hours are regulated (eg, 80-hour resident special act in Korea and The European Working Time Directive in Europe) [18,20]. Of course, there may be some opinions concerning undertraining of residents, but the introduction of educational tools, such as simulation training systems, is a possible alternative [17,21].

Issues Related to Practical Application

Before adopting and introducing an AI-based surgical assistant system in the surgical field, ethical and legal responsibilities should be discussed through consensus of medical, legal, and administrative experts and others. Additionally, although not included in this survey, the recent development of AI is likely to include explainable AI, a concept contrasted with previous black-box AI, in the development of new technologies.

At the time of the introduction of robotic surgery, which is being actively used presently, many experts had discussed ethical issues [22-24]. Current robotic surgery is a master-slave system, with the surgeon having most of the responsibility, making it easy to discuss ethical issues. However, in the case of an autonomous AI-based surgical assistant system, there may be controversy regarding the responsibility for harm and injuries caused to the patient during the robotic surgery, and social discussions about this need to be carried out for the adoption of an AI-based surgical assistant system [24,25].

Explainability should be considered when newly developing AI-based surgical assistant systems. Current AI-based medical programs involving deep learning and machine learning techniques lack explainability, hindering the dependence of medical professionals on conclusions from these programs. Therefore, considering the characteristics of surgical procedures that are repeated continuously with small and large decisions, it is expected that explainability will be essential for the interaction between the machine and the operator and should be incorporated in the development of AI-based robotic assistance systems that contribute to these procedures [26].

Strengths and Weaknesses

This report is based on a survey among experts who have been actively performing laparoscopic surgery in various fields. To the best of our knowledge, this is the first report showing the clinical need for an AI-based surgical assistant system for AISS according to an e-survey. Moreover, this study is meaningful because we could identify the unmet need of clinicians for an AI-based system for AISS, which could be developed soon. However, this study has some limitations. First, it was challenging to check the exact response rate through the mailing system used in this study, which could act as a bias, and thus, the results of this study should be interpreted carefully. However, we could assume that the questionnaire was answered by our targeted responders because most of the responders mentioned that they performed more than one surgery per month. Second, the specific national health insurance system controlled by the government in the Republic of Korea could affect the expected value of an AI-based surgical assistant system for AISS, and the finding should be complemented by international surveys later. Third, the validity and reliability of the items in the questionnaire could not be confirmed because there has been no previous comparable study and this study targeted a specific group of experts in our country.

Conclusion

In the conventional laparoscopic surgical system, surgeons may value the proficiency of assistants, and most of them may feel uncomfortable with the unintended or not intuitive movement of laparoscopic cameras and devices. For the development of an AI-based surgical assistant system in the future, safe operation may be expected through lens cleaning, intuitive manipulation, and tissue damage minimization. Furthermore, an AI-based surgical assistant system is expected to replace approximately 41%-60% of the workforce, which may increase surgeons' willingness to purchase such a system for reducing human resources and performing a comfortable, safe, and skilled operation. Conclusively, an AI-based surgical assistant system for AISS will become essential to enhance surgeons' convenience, but it will be necessary to increase the safety and quality of surgery for patients.

Acknowledgments

This study was presented at the 27th Annual Meeting of the Korean Society of Gynecologic Endoscopy and Minimal Invasive Surgery and received the Best Oral Presentation Award. This research was supported by grants from Seoul National University (800-20170249, 800-20180201, and 800-20190437) and the National Research Foundation of Korea (2017057240).

Authors' Contributions

HSK designed this survey; HSK, EJL, SIK, SHK, and CWJ collected the data; and SJP performed the analysis and wrote the manuscript. All authors contributed to the revision of the manuscript.

Conflicts of Interest

None declared.

References

1. Miller DD, Brown EW. Artificial Intelligence in Medical Practice: The Question to the Answer? Am J Med 2018 Feb;131(2):129-133. [doi: 10.1016/j.amjmed.2017.10.035] [Medline: 29126825]
2. Ramesh A, Kambhampati C, Monson J, Drew P. Artificial intelligence in medicine. Ann R Coll Surg Engl 2004 Sep 01;86(5):334-338 [FREE Full text] [doi: 10.1308/147870804290] [Medline: 15333167]
3. Mirnezami R, Ahmed A. Surgery 3.0, artificial intelligence and the next-generation surgeon. Br J Surg 2018 Apr;105(5):463-465. [doi: 10.1002/bjs.10860] [Medline: 29603133]
4. Perez RE, Schwaitzberg SD. Robotic surgery: finding value in 2019 and beyond. Ann. Laparosc. Endosc. Surg 2019 May;4:51-51. [doi: 10.21037/ales.2019.05.02]
5. El Hachem L, Momeni M, Friedman K, Moshier EL, Chuang LT, Gretz HF. Safety, feasibility and learning curve of robotic single-site surgery in gynecology. Int J Med Robot 2016 Sep 11;12(3):509-516. [doi: 10.1002/rcs.1675] [Medline: 26096813]
6. Thiel DD, Winfield HN. Robotics in urology: past, present, and future. J Endourol 2008 Apr;22(4):825-830. [doi: 10.1089/end.2007.9830] [Medline: 18419224]
7. Sutherland GR, Wolfsberger S, Lama S, Zarei-nia K. The Evolution of neuroArm. Neurosurgery 2013;72:A27-A32. [doi: 10.1227/neu.0b013e318270da19]
8. van den Berg J, Miller S, Duckworth D, Hu H, Wan A. Superhuman performance of surgical tasks by robots using iterative learning from human-guided demonstrations. 2010 Presented at: 2010 IEEE International Conference on Robotics and Automation; May 3-8, 2010; Anchorage, Alaska. [doi: 10.1109/robot.2010.5509621]
9. Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC. Supervised autonomous robotic soft tissue surgery. Sci Transl Med 2016 May 04;8(337):337ra64-337ra64. [doi: 10.1126/scitranslmed.aad9398] [Medline: 27147588]
10. Miehle J, Ostler D, Gerstenlauer N, Minker W. The next step: intelligent digital assistance for clinical operating rooms. Innov Surg Sci 2017 Sep;2(3):159-161 [FREE Full text] [doi: 10.1515/iss-2017-0034] [Medline: 31579748]
11. Mewes A, Hensen B, Wacker F, Hansen C. Touchless interaction with software in interventional radiology and surgery: a systematic literature review. Int J Comput Assist Radiol Surg 2017 Feb 19;12(2):291-305. [doi: 10.1007/s11548-016-1480-6] [Medline: 27647327]
12. Mentis H, O'Hara K, Gonzalez G, Sellen A, Corish R, Criminisi A. Voice or Gesture in the Operating Room. 2015 Apr Presented at: CHI '15: CHI Conference on Human Factors in Computing Systems; April, 2015; Seoul Republic of Korea. [doi: 10.1145/2702613.2702963]
13. Barbash GI, Glied SA. New Technology and Health Care Costs — The Case of Robot-Assisted Surgery. N Engl J Med 2010 Aug 19;363(8):701-704. [doi: 10.1056/nejmp1006602]
14. Kalantari DH, Kalantari DE, Maleki S. E-survey (surveys based on e-mail & web). Procedia Computer Science 2011;3:935-941. [doi: 10.1016/j.procs.2010.12.153]
15. nownsurvey.: ELIMNET Co, Ltd URL: https://www.nownsurvey.com/ [accessed 2020-04-03]
16. van den Haak L, Alleblas C, Nieboer TE, Rhemrev JP, Jansen FW. Efficacy and safety of uterine manipulators in laparoscopic surgery: a review. Arch Gynecol Obstet 2015 Nov 13;292(5):1003-1011. [doi: 10.1007/s00404-015-3727-9] [Medline: 25967852]
17. Kauvar DS, Braswell A, Brown BD, Harnisch M. Influence of resident and attending surgeon seniority on operative performance in laparoscopic cholecystectomy. J Surg Res 2006 May 15;132(2):159-163. [doi: 10.1016/j.jss.2005.11.578] [Medline: 16412471]
18. Kim DJ, Kim SG. Comparative study of the operative experience of surgical residents before and after 80-hour work week restrictions. Ann Surg Treat Res 2018 Nov;95(5):233-239 [FREE Full text] [doi: 10.4174/astr.2018.95.5.233] [Medline: 30402441]
19. Coverdill JE, Shelton JS, Alseidi A, Borgstrom DC, Dent DL, Dumire R, et al. The promise and problems of non-physician practitioners in general surgery education: Results of a multi-center, mixed-methods study of faculty. Am J Surg 2018 Feb;215(2):222-226. [doi: 10.1016/j.amjsurg.2017.10.040] [Medline: 29137723]
20. Fitzgerald J, Caesar B. The European Working Time Directive: a practical review for surgical trainees. Int J Surg 2012;10(8):399-403 [FREE Full text] [doi: 10.1016/j.ijsu.2012.08.007] [Medline: 22925631]
21. Graafland M, Bok K, Schreuder HW, Schijven MP. A multicenter prospective cohort study on camera navigation training for key user groups in minimally invasive surgery. Surg Innov 2014 Jun 16;21(3):312-319. [doi: 10.1177/1553350613505714] [Medline: 24132469]
22. Larson JA, Johnson MH, Bhayani SB. Application of surgical safety standards to robotic surgery: five principles of ethics for nonmaleficence. J Am Coll Surg 2014 Feb;218(2):290-293. [doi: 10.1016/j.jamcollsurg.2013.11.006] [Medline: 24315652]
23. Siqueira-Batista R, Souza CR, Maia PM, Siqueira SL. Robotic Surgery: Bioethical Aspects. Arq Bras Cir Dig 2016 Dec;29(4):287-290 [FREE Full text] [doi: 10.1590/0102-6720201600040018] [Medline: 28076489]

24. O'Sullivan S, Nevejans N, Allen C, Blyth A, Leonard S, Pagallo U, et al. Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. Int J Med Robot 2019 Feb 09;15(1):e1968. [doi: 10.1002/rcs.1968] [Medline: 30397993]
25. O'Sullivan S, Leonard S, Holzinger A, Allen C, Battaglia F, Nevejans N, et al. Anatomy 101 for AI-driven robotics: Explanatory, ethical and legal frameworks for development of cadaveric skills training standards in autonomous robotic surgery/autopsy. Int J Med Robot 2019 May 30:e2020. [doi: 10.1002/rcs.2020] [Medline: 31144777]
26. Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 2019 Apr 02;9(4):e1312 [FREE Full text] [doi: 10.1002/widm.1312] [Medline: 32089788]

Abbreviations

AI: artificial intelligence
AISS: artificial intelligence–based solo surgery

Edited by G Eysenbach; submitted 31.12.19; peer-reviewed by A Holzinger, N Mohammadzadeh; comments to author 25.02.20; revised version received 10.03.20; accepted 12.03.20; published 15.05.20.

Please cite as:
Park SJ, Lee EJ, Kim SI, Kong SH, Jeong CW, Kim HS
Clinical Desire for an Artificial Intelligence–Based Surgical Assistant System: Electronic Survey–Based Study
JMIR Med Inform 2020;8(5):e17647
URL: http://medinform.jmir.org/2020/5/e17647/
doi:10.2196/17647
PMID:32412421

Original Paper

The Development of the Military Service Identification Tool: Identifying Military Veterans in a Clinical Research Database Using Natural Language Processing and Machine Learning

Daniel Leightley[1], BSc, MSc, PhD; David Pernet[1], BA; Sumithra Velupillai[2,3], MA, PhD; Robert J Stewart[2,3], FRCPsych; Katharine M Mark[1], BSc, MSc, PhD; Elena Opie[1], MSc; Dominic Murphy[1,4], MA, PhD, DClinPsy; Nicola T Fear[1,5], BSc, MSc, DPhil; Sharon A M Stevelink[1,6], BSc, MSc, PhD

[1]King's Centre for Military Health Research, King's College London, London, United Kingdom

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

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

[4]Combat Stress, Letherhead, United Kingdom

[5]Academic Department of Military Mental Health, King's College London, London, United Kingdom

[6]Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

Corresponding Author:
Daniel Leightley, BSc, MSc, PhD
King's Centre for Military Health Research
King's College London
London
United Kingdom
Phone: 44 20 7848 5351
Email: daniel.leightley@kcl.ac.uk

Abstract

Background: Electronic health care records (EHRs) are a rich source of health-related information, with potential for secondary research use. In the United Kingdom, there is no national marker for identifying those who have previously served in the Armed Forces, making analysis of the health and well-being of veterans using EHRs difficult.

Objective: This study aimed to develop a tool to identify veterans from free-text clinical documents recorded in a psychiatric EHR database.

Methods: Veterans were manually identified using the South London and Maudsley (SLaM) Biomedical Research Centre Clinical Record Interactive Search—a database holding secondary mental health care electronic records for the SLaM National Health Service Foundation Trust. An iterative approach was taken; first, a structured query language (SQL) method was developed, which was then refined using natural language processing and machine learning to create the Military Service Identification Tool (MSIT) to identify if a patient was a civilian or veteran. Performance, defined as correct classification of veterans compared with incorrect classification, was measured using positive predictive value, negative predictive value, sensitivity, F1 score, and accuracy (otherwise termed Youden Index).

Results: A gold standard dataset of 6672 free-text clinical documents was manually annotated by human coders. Of these documents, 66.00% (4470/6672) were then used to train the SQL and MSIT approaches and 34.00% (2202/6672) were used for testing the approaches. To develop the MSIT, an iterative 2-stage approach was undertaken. In the first stage, an SQL method was developed to identify veterans using a keyword rule–based approach. This approach obtained an accuracy of 0.93 in correctly predicting civilians and veterans, a positive predictive value of 0.81, a sensitivity of 0.75, and a negative predictive value of 0.95. This method informed the second stage, which was the development of the MSIT using machine learning, which, when tested, obtained an accuracy of 0.97, a positive predictive value of 0.90, a sensitivity of 0.91, and a negative predictive value of 0.98.

Conclusions: The MSIT has the potential to be used in identifying veterans in the United Kingdom from free-text clinical documents, providing new and unique insights into the health and well-being of this population and their use of mental health care services.

(JMIR Med Inform 2020;8(5):e15852) doi:10.2196/15852

KEYWORDS

natural language processing; machine learning; military personnel; electronic health care records; mental health; veteran

Introduction

Veterans

Estimates of the United Kingdom's military veteran population, defined by the British Government as those who have served in the military for at least one day [1], are approximately 2.5 million, equivalent to approximately 5% of household residents aged 16 years or older in the United Kingdom [2]. UK military veterans receive health care provision from the National Health Service (NHS) alongside civilians, with care recorded in local, regional, and national electronic health care records (EHRs) [3]. EHRs—structured and unstructured (ie, free text)—can be used to evaluate disease prevalence and surveillance, to perform epidemiological analyses and investigate the quality of care, and to improve clinical decision making [4,5].

Veterans of the United Kingdom experience a range of mental health problems (estimates range from 7% to 22% across psychiatric conditions), some resulting from their experiences in the line of duty [6]. A large UK cohort study set up to investigate the health of serving personnel and veterans has also shown that veterans report higher levels of probable posttraumatic stress disorder and alcohol misuse than serving personnel [7]. Recent research suggests that 93% of veterans who report having a mental health difficulty seek some form of help for their problems, including informal support through family and friends [8]. However, there is no national marker in the UK EHRs to identify veterans, nor is there a requirement for health care professionals to record it, making it difficult to evaluate the unique health care needs of those who have served in the UK Armed Forces [9]. Furthermore, the ability to identify veterans would allow for comparisons between civilian and military cohorts and for direct comparison of their physical and mental health.

In England and Wales, only two studies exist, which analyze secondary care delivered through the NHS for Armed Forces personnel. In the first study, Leightley et al [3] developed a method to link the EHRs of military personnel in England, Scotland, and Wales (3 nations of the United Kingdom). This study used a longitudinal cohort consisting of serving personnel and veterans to establish a link to national EHRs (England, Scotland, and Wales). Then, statistical analyses were performed to identify the most common reasons for admission into hospital, diagnoses, and treatment pathways. The second study, by Mark et al [10], on which this study is based, systematically searched for veterans using a military-related search term strategy on free-text clinical documents using a manual approach. Although this approach could identify veterans, it was time consuming as searches were performed manually. Each of these studies highlighted a need for novel methodological development for the identification of veterans, with natural language processing (NLP) and machine learning showing great promise [11-13]. This would enable the automatic identification of veterans without the need for manual annotation and validation.

Neutral Language Processing

NLP approaches cover wide-ranging solutions to the analysis of text, such as retrieval, analysis, transformation, and classification of text, such as those found in EHRs and free-text clinical documents [13,14]. NLP subthemes, such as text mining, are represented as a set of programmatic rules or machine learning algorithms (eg, automated learning from labeled data) to extract meaning from naturally occurring text (eg, human-generated text) [11,14]. The result is often an output that can be interpreted by humans and that can be processed computationally more efficiently [15]. It may be possible to apply NLP for the identification of veterans, if not already defined from structured fields, such as a flag for denoting veteran status, for which, in the United Kingdom, are rarely coded [10]. The ability to identify veterans at scale could significantly improve our understanding of their health and well-being and navigation of care pathways and allow for the exploration of the long-term impacts of service.

This Study

NLP tools have been used extensively in military health research, predominantly in the United States, for the detection of veteran homelessness and clinical diagnosis [16-19]. However, to the best of our knowledge, no tools exist to identify veteran status using either a rule-based or machine learning approach. The aim of this study was to describe the development of the Military Service Identification Tool (MSIT) for the identification of veterans using free-text clinical documents and to evaluate the tool's performance against a manually annotated dataset (gold standard). This study was inspired by the study by Fernandes et al [14], but we proposed a different approach to the way in which features are generated and used for training machine learning classifiers and the annotation of the training and testing data and the way in which we evaluate the performance of MSIT across different classifiers.

Methods

Data Source—Clinical Record Interactive Search System

The Clinical Record Interactive Search (CRIS) system provides deidentified EHRs from the South London and Maudsley (SLaM) NHS Foundation Trust, a secondary and tertiary mental health care provider serving a geographical catchment of approximately 1.3 million residents of 4 south London boroughs (Lambeth, Southwark, Lewisham, and Croydon) [20]. The CRIS system has supported a range of research projects [20-23]. Many of these have aimed to answer specific clinical or epidemiological research questions and have drawn on particular subpopulations being identified in the database, such as ethnic minorities and those with Alzheimer disease [24,25].

Ethical approval for the use of CRIS as an anonymized database for secondary analysis was granted by the Oxford Research Ethics Committee (reference: 08/H0606/71+5). This study has been approved by the CRIS Patient Data Oversight Committee

of the National Institute of Health Research (NIHR) Biomedical Research Centre (reference: 16-056).

The documents used in this study are Correspondence, which are created by clinical staff to provide a summary of admission or care received and are sent to a patient's general practitioner and, in some cases, to the patient themselves. Correspondence were used as they routinely provided a detailed history of a patient's life events including employment history.

Study Design

There are approximately 300,000 correspondence documents available in CRIS. Owing to the large volumes of data, a subset was extracted for the development of the MSIT. This subset (hereafter termed personal history dataset) was extracted using the personal history detection tool, which has been developed by the CRIS team [26]. This tool identifies documents that have a subheading or section entitled personal history (or similar) before extracting the proceeding text (see Textbox 1 for an example). Each personal history record contains an outline of each patient's life events since birth (eg, educational attainment, childhood adversity, employment, and relationship information). Each record is written by a clinician. The personal history dataset contains 98,395 documents sampled from records recorded in CRIS since 2006, which was the first year the CRIS database was operational.

After an informal scoping exercise, discussions with NLP experts with experience of using CRIS and timing constraints of the study, the decision was made to retain only 6672 documents (hereafter termed gold standard dataset), which represented 4200 patients (civilian: 3331 and veteran: 869). A patient could have multiple documents that represent different time points of care. The decision to retain 4200 patients (which in total had 6672 documents) was made considering resource limitations of the study, which included staff time to annotation and balancing patient privacy as to only process a minimum number of records to allow us to archive the study aim. A sample size calculation was not performed because of these considerations.

For evaluating the performance of MSIT, a decision was made to retain 66.00% (4470/6672 documents) of the dataset for training, and the remainder 34.00% (2202/6672 documents) was used for testing and evaluation. Patients were sampled to either the training or testing; dataset a patient's documents would not appear in both samples. There is no defined approach for determining the size of the training and testing sets needed, with most research using ad hoc reasoning depending on data, financial, time, or personal constraints [27]. This study followed an iterative approach to the development of the MSIT, first by developing a structured query language (SQL) rule–based method, with lessoned learned, such as which keywords cause misclassification, informing the development of MSIT.

Textbox 1. Synthetically generated personal history statement by the research team for a female patient whose father and husband served in the military. X denotes personal identifier being removed. Owing to patient confidentiality, we were not able to share real examples from the personal history dataset.

> *Mrs X was born in X. Her father was a Normandy D-Day veteran who had sustained a bullet wound to his left arm during the war. He subsequently worked as a bus driver in and around X. Mrs X describes her upbringing as old-fashioned, traditional and one of poverty. She describes her school years as happy and fun and says she got on well with her parents. She acknowledged that during her teenage years that she was difficult to manage. She met her husband X while on holiday in X; X was stationed there in a military unit conducting NATO exercises. After they began a relationship, in 1983, they moved to X. Mrs X worked in various jobs including in a supermarket and as a hotel receptionist, before taking an administrative job in academia.*

Generating the Gold Standard Dataset and Interrater Agreement

A set of classification rules for the annotation of each document was developed and agreed upon by DL, EO, DP, and SS. The Extensible Human Oracle Suite of Tools (University of Utah) software package was used to perform annotations [28]. The following words and phrases were annotated: (1) those that described a patient's military service (ie, "he served in the Army"), (2) those that described an individual other than the patient's military service (ie, "dad served in the Forces"), and (3) those that may cause confusion (ie, "Navy Blue"). This led to the creation of a gold standard dataset, which contained veterans- and civilians-annotated free-text clinical documents. Veterans were labeled as such based on a clear statement that the patients themselves had served in the military. The protocol, including classification rules, is available on request from the corresponding author.

Developing a Rule-Based Approach for Veteran Identification

Civilians and veterans were classified using the SQL rule–based method based on a corpus of known words and phrases related to military service (see Multimedia Appendix 1). The corpus was composed of (1) primary search terms: common words or phrases used to describe military service, (2) secondary search terms: used to validate that the document describes a patient who has served in the military, and (3) exclusion terms: used to exclude documents that may describe another person's military service and not the patient's military service.

The SQL rule–based method was developed using a combination of the research team's expert knowledge of the military, relevant research literature, and analysis of personal history statements. The gold standard training dataset was used to refine the SQL rule–based approach. The code was iteratively tested on the training set, reviewed, and refined to ensure full coverage of known military words and phrases. The SQL rule–based method operated by searching for the occurrence of a primary search term in a document. If the term was found, text surrounding the

term would be extracted (up to 50 characters, where available). The extracted text was then evaluated against a list of secondary terms to classify the document as a civilian document or a veteran document. The SQL rule–based approach informed the development of the MSIT.

Developing the Military Service Identification Tool

A machine learning classification framework was used to create MSIT. It was developed in Python using the Natural Language Processing Toolkit (version 3.2.5) [29] and Scikit-learn (version 0.20.3) [30]. The gold standard dataset was preprocessed to remove (1) punctuations (using regular expressions), (2) words/phrases related to another individual's military service (these were required to exactly match those in the gold standard annotated dataset), (3) stop words and frequently occurring terms (except military terms), and (4) word/phrases that may cause confusion with correctly identifying a veteran. The remaining features were then converted into term frequency-inverse document frequency (tf-idf) features.

The classification framework was trained to identify veterans based on the use of military terms and phrases with the outcome being binary (1: veteran and 0: not a veteran). A training set of 4470 annotated documents was used to select a machine learning classifier. There is sparse literature on which machine learning algorithms are best suited for specific tasks, not only in the field of NLP but also in areas such as health care, agriculture, and security [31-34]. To ensure the appropriate selection of the classifier used for the MSIT, a comparison was made based on 10-fold cross-validation accuracy using tf-idf features as an input of the following machine learning classifiers (which are part of the Scikit-learn package): random forest, decision tree, linear support vector classifier, support vector classifier, multinomial Naïve Bayes, k-nearest neighbor, logistic regression, and multilayered perception. Each machine learning classifier used default parameters. Linear support vector classifier obtained the highest accuracy (see Table 1; accuracy=0.95; SD 0.01; 95% CI 0.94-0.95) and was used as the machine learning classifier for MSIT.

To improve the true positive rate of the MSIT and to reduce the potential for false positives, a postprocessing of the linear support vector classifier outcome was applied based on the SQL rule–based approach described earlier, as has been used in similar studies [14]. For each document that was predicted as being that of a veteran, an SQL operation was performed to ensure the document used a military term of phrase (eg, "joined the army," "left the army," and "demobbed from the army").

Table 1. Machine learning classifier n-fold cross-validation accuracy, SD, and 95% CI based on the gold standard training dataset of 4470 documents.

Classifier	Accuracy	SD	95% CI
Random forest	0.84	0.01	0.83-0.84
Decision tree	0.91	0.03	0.89-0.92
Linear support vector classifier	0.95	0.01	0.94-0.95
Support vector classifier	0.84	0.01	0.83-0.84
Multinomial Naïve Bayes	0.90	0.02	0.88-0.91
k-nearest neighbor	0.89	0.02	0.87-0.90
Logistic regression	0.88	0.04	0.85-0.90
Multilayered perception	0.94	0.02	0.92-0.95

Availability of Materials and Data

The datasets used in this study are based on patient data, which are not publicly available. Although the data are pseudonymized, that is, personal details of the patient are removed, the data still contain information that could be used to identify a patient. Access to these data requires a formal application to the CRIS Patient Data Oversight Committee of the NIHR Biomedical Research Centre. On request and after suitable arrangements are put in place, the data and modeling employed in this study can be viewed within the secure system firewall. The corresponding author can provide more information about the process.

A Jupyter Notebook demonstrating the tool with artificial data can be found in the link provided [35].

Statistical Analyses

All analyses were performed using Python version 3.5 with standard mathematical packages and Scikit-learn (version 0.20.3) [30]. Cohen kappa values are presented for civilian and veteran annotations separately, with a two-tailed statistical test applied to determine the significance of the finding. Machine learning classifier 10-fold cross-validation was reported as the highest accuracy obtained, with SD and 95% CI reported to represent the n-fold result. Document characteristics were reported as the average frequency in which words, sentences, whitespaces, stop words, and nonalphanumeric across documents were stratified by civilian and veteran. The most frequent military terms and phrases annotated during the study were restricted to the top 5 and reported as a count with percentage out of the denominator. For evaluating the SQL rule–based approach, the algorithm was tested by measuring the output results against the results from manual annotations (the gold standard testing dataset), allowing for computation of positive predictive value, negative predictive value, sensitivity, F1 score, and accuracy at a document level. For evaluating MSIT, each classifier model was tested by measuring its results against the results from manual annotations (the gold standard testing dataset), allowing for computation of positive predictive value, negative predictive value, sensitivity, F1 score, and accuracy at a document level.

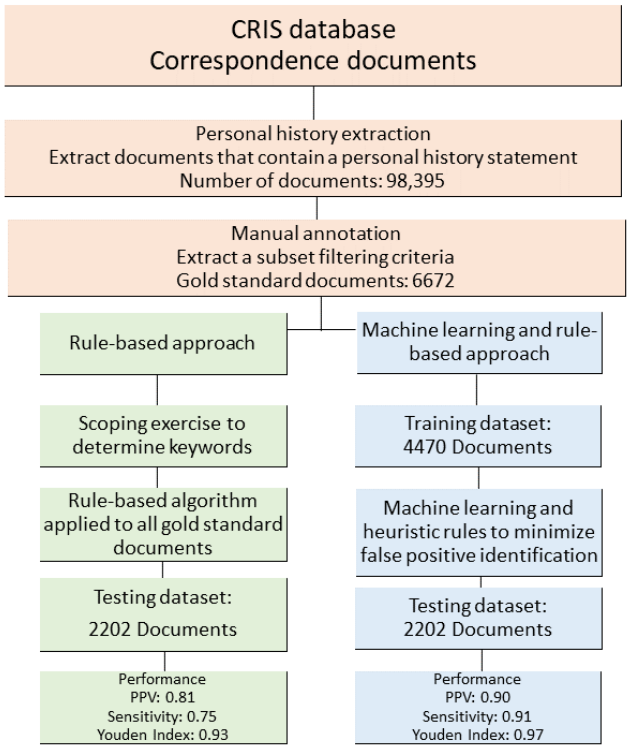
In this study, positive predictive value was defined as the proportion of correctly identified true veterans over the total number of true veterans identified by the classifier. Negative predictive value was defined as the proportion of correctly identified true civilians over the total number of true civilians identified by the classifier. Sensitivity was defined as the proportion of true veterans identified by the classifier over the total number of actual veterans (identified by manual annotation). F1 score considers both positive predictive value and sensitivity and produces a harmonic mean, where the best value lies at 1 and the worst value lies at 0. Accuracy was measured using the Youden Index, which considers sensitivity and specificity (summation minus 1), which results in a value that lies between 0 (absence of accuracy) and 1 (perfect accuracy).

Results

Annotation

An iterative approach to developing MSIT was employed. See Figure 1 for a flow diagram of the MSIT and evaluation process. The datasets used in this study were independently annotated by DL, EO, and a researcher (see Acknowledgments section), with acceptable interrater agreement as indicated by a Cohen kappa of 0.83 for veterans and 0.89 for civilians (P=.15).

Figure 1. Flow diagram of the Military Service Identification Tool. Correspondences are used to define any communications between a patient and clinical staff or between clinical staff members. CRIS: Clinical Record Interactive Search; PPV: Positive Predictive Value.



Document Characteristics

Of the 6672 documents annotated to generate the gold standard dataset, there were 5630 civilian and 1042 veteran documents. Descriptive characteristics (see Table 2) indicate that often civilian documents had more words, sentences, stop words, and nonalphanumeric characters.

A total of 2611 words and 2016 phrases that describe a patient's military service were annotated (see Tables 3 and 4). Most of the words and phrases annotated described the service branch (eg, "served in the army," "national service in the RAF," "demobbed from the army," and "was a pilot in the RAF"), with only a small number including the length of service (eg, "served for two years in the army," "served two years for national service," and "demobbed from the army after two years").

Table 2. Document characteristics including frequency and mean (SD) for annotated personal history statements stratified by civilian and veteran status.

Characteristic	Civilian (n=5630), mean (SD)	Veteran (n=1042), mean (SD)
Words	223.76 (152.30)	197.20 (114.63)
Sentences	13.80 (8.91)	12.40 (6.50)
Whitespaces	237.99 (162.77)	208.38 (119.65)
Stop words	32.04 (11.45)	30.09 (9.92)
Nonalphanumeric characters	26.59 (20.14)	22.22 (14.28)

Table 3. Top 5 occurring military words identified during manual annotation of the gold standard training dataset.

Military words (n=2611)	Value, n (%)
“Army”	553 (21.18)
“National Service”	445 (17.04)
“RAF”	225 (8.62)
“Navy”	166 (6.36)
“Veteran”	104 (4.00)

Table 4. Top 5 occurring military phrases identified during manual annotation of the gold standard training dataset.

Military phrases (n=2016)	Value, n (%)
“Joined the army”	167 (8.28)
“Left the army”	122 (6.05)
“Demobbed from the army”	101 (5.00)
“National service in the army”	65 (3.22)
“Two years in the army”	64 (3.22)

Performance: Positive Predictive Value, Sensitivity, and Accuracy

The performance of each approach was evaluated against the manually annotated gold standard test dataset producing positive predictive value, negative predictive value, sensitivity, F1 score, and accuracy statistics. The gold standard test dataset contained 2202 documents, which included 1882 civilian and 320 veteran documents (see Tables 5 and 6).

The SQL rule–based approach correctly identified 262 veteran documents, incorrectly identified 87 civilian documents as veteran documents, and incorrectly identified 58 veteran documents as civilian documents. Misclassification was because of the rigidity of the keywords used to search the records, with confusion observed between the individual’s serving status and a family member’s status. For example, phrases such as “had served” were used to describe another person’s military service, such as father or brother. This resulted in an overall accuracy of 0.93, a positive predictive value of 0.81, a negative predictive value score of 0.95, a sensitivity of 0.75, and an F1 score of 0.78.

During the initial development of the MSIT, model sensitivity was skewed toward commonly occurring words. To overcome this bias, a 4-step preprocessing step was introduced to identify and remove these frequent words and phrases, punctuation, and stop words, which improved positive predictive value and sensitivity of the tool (training dataset: positive predictive value=0.78 and sensitivity=0.88). To further improve the prediction of the tool and reduce the potential for false positives, a postprocessing step was introduced to ensure a military word or phrase was present in the documents predicted as describing a veteran. The addition of this step improved positive predictive value and sensitivity of the MSIT (training dataset: positive predictive value=0.82 and sensitivity=0.91).

Applying MSIT to the gold standard test dataset correctly identified 290 veteran documents, incorrectly identified 30 civilian documents as veteran documents, and incorrectly identified 27 civilian documents as being a veteran document. Misclassification was observed, with manual inspection of the documents revealing that the military-related terms were used to describe events, occupations, or items for civilians such as “Legion” or “Mess Hall.” This created confusion with the classifier. This resulted in an overall accuracy of 0.97, a positive predictive value of 0.90, a negative predictive value of 0.95, a sensitivity of 0.91, and an F1 score of 0.91. Additional analyses were conducted using the leave-one-out methodology (see Multimedia Appendix 1).

Table 5. Confusion matrix indicating the performance of the structured query language rule-based approach and the Military Service Identification Tool (MSIT). The MSIT includes pre- and postprocessing.

Label	Structured query language rule–based approach		Military Service Identification Tool	
	Veteran	Civilian	Veteran	Civilian
Veteran	262	58	290	30
Civilian	87	1795	27	1855

Table 6. Structured query language–based approach and Military Service Identification Tool (MSIT) performance result comparison for detecting veterans using the gold standard test dataset. The MSIT includes pre- and postprocessing.

Performance metric	Structured query language rule–based approach	Military Service Identification Tool
Positive predictive value	0.81	0.90
Negative predictive value	0.95	0.98
Sensitivity	0.75	0.91
F1 score	0.78	0.91
Youden Index	0.93	0.97

Discussion

Principal Findings

This research has demonstrated that it is possible to identify veterans from free-text clinical documents using NLP. A tool to identify veterans and civilians is described, which performed well, as indicated by high positive predictive value, sensitivity, and accuracy results. To the authors' knowledge, this is the only study to have developed, applied, and tested NLP for the identification of veterans in the United Kingdom using a large psychiatric database. The MSIT presented superior results to the SQL rule–based approach developed because of the former's ability to adapt to different military terms. The SQL rule–based approach was, on the other hand, fixed on set keywords.

This is the first study that seeks to identify military veterans from a case register in the United Kingdom using NLP and machine learning. Although military literature is sparse, NLP techniques have been used in the detection of sexual trauma, in the detection of temporal expressions in medical narratives, and for screening homelessness [16,17,19]. Although it is difficult to compare our study with the aforementioned studies, similar methodologies are employed. This includes each developing a gold standard manually annotated dataset, developing a set of rules to support identification, and finally generating features from free text. Although this study used linear support vector classification, as it was determined to be the most optimal, Reeves et al [16] used a maximum entropy classifier to detect temporal expressions. Outside of the military literature, Fernandes et al [14] sought to identify suicidal attempts using a psychiatric database with support vector machines; they were able to detect suicidal attempts with a sensitivity of 0.98, which is higher than what was achieved in this study (MSIT: 0.91). Other studies have compared different classification algorithms for clinical NLP tasks with varying conclusions—achieving optimal performance is highly task-dependent and use-case–dependent [36,37].

The ability to identify veterans could provide insights into the physical and mental health of military personnel and their navigation through, and use of, health care services, including primary and secondary services. This would overcome the current need to either manually identify veterans or to perform large-scale cohort and data linkage studies, such as that by Leightley et al [3]. EHR-based case registers, such as CRIS, function as single, complete, and integrated electronic versions of traditional paper health records [3]. These registers have been positioned as a new generation for health research and are now mandatory in the United Kingdom [3]. The methodological advantages of case registers—including their longitudinal nature, largely structured fields, and detailed coverage of defined populations—make them an ideal research and surveillance tool [38]. EHRs in mental health care provide extremely rich material, and analysis of their data can reveal patterns in health care provisions, patient profiles, and mental and physical health problems [3,39]. EHRs are advantageous for investigating vulnerable subgroups within the wider population [20-22], potential for developing digital interventions [40] and to support data-driven decision making [11].

Strengths and Limitations

An important strength of this work was the exploitation of NLP, which is advantageous for automating the process of identification and reducing the possibility of human error and bias. Considering the focus of this study, this is the first time that NLP has successfully been used to identify veterans from free-text clinical documents using detailed occupational history that clinicals routinely record. The MSIT described in this work does not rely on any codes (clinical or otherwise) or structured fields, which broadens its application to others, such as diagnosis and occupation detection. Furthermore, veterans may not always be willing or think it is necessary to state their veteran status, particularly in the United Kingdom, which has no department for veterans' affairs. As such, NLP is advantageous, as it may pick up veterans based on small details that are discussed and recorded during clinical interactions rather than having to rely on disclose of veteran status by an individual upon registration with clinical services.

It must be noted that there are several limitations to the tool described in this work. First, the study relied on patients' self-reporting that they have served in the military, which could be influenced by the patient's mental health or failing memory. Second, the need for a clinician to ask a patient's military status and for this to be accurately recorded in the patient notes. Third, the accuracy of recording by the clinician could have had a negative impact on MSIT's performance or could result in misidentification of veterans. Fourth, the MSIT relied on the personal history section being present in a correspondence, which may limit scalability. Fifth, although different approaches to stating veteran service were annotated, spelling and additional permutations were not considered. This could limit the generalizability of the algorithms on other datasets. Sixth, identified veterans were not validated against the Ministry of Defence databases or contacted directly to validate veteran status. Seventh, a sample size calculation was not computed for this study. This was because of resource limitations; as a result,

this could limit the generalizability of the algorithms on other datasets. Finally, documents were misclassified, often because of military vernacular being used by civilians and/or the clinician or because a family member had served in the military and not the patient. Further work should be undertaken to improve reliability and reduce the rate of misclassification.

Conclusions

We have shown that it is possible to identify veterans using either an SQL-based or NLP- and machine learning–based approach. Both approaches are robust in correctly identifying civilians and veterans, with high accuracy, sensitivity, and negative predictive values observed. The MSIT has the potential to be used in identifying veterans in the United Kingdom from free-text clinical documents, providing new and unique insights into the health and well-being of this population and their use of mental health care services. Despite our success in this work, the tools are tailored to the CRIS dataset, and future work is needed to develop a more agnostic framework.

Acknowledgments

This study was funded by the Forces in Mind Trust (Project: FiMT18/0525KCL), a funding scheme run by the Forces in Mind Trust using an endowment awarded by the National Lottery Community Fund. The salary of SV, RS, and SS was partly paid by the NIHR Biomedical Research Centre at the SLaM NHS Foundation Trust and King's College London. In addition to the listed authors, the study involved support from the NIHR Biomedical Research Centre. NIHR Biomedical Research Centre is a partnership between the SLaM NHS Foundation Trust and the Institute of Psychiatry, Psychology, and Neuroscience at King's College London. The authors would particularly like to thank Megan Pritchard (lead in CRIS training and development), Debbie Cummings (administrator), Karen Birnie (researcher), and Larisa Maria (researcher) for their help and support in undertaking this study.

Authors' Contributions

SM, DM, and NF conceived the concept of the study and obtained funding. DL and DP developed the NLP approaches used in this study. DL, KM, and EO performed data annotation. SV and RS provided substantial improvements to the manuscript after drafting. All authors reviewed the final manuscript.

Conflicts of Interest

NF, DP, and SS are partly funded by the United Kingdom's Ministry of Defence. NF sits on the Independent Group Advising on the Release of Data at NHS Digital. NF is also a trustee of two military-related charities. DM is employed by Combat Stress, a national charity in the United Kingdom that provides clinical mental health services to veterans. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health and Social Care, or the UK Ministry of Defence.

Multimedia Appendix 1
Supplementary material.
[PDF File (Adobe PDF File), 571 KB - medinform_v8i5e15852_app1.pdf]

References

1. Ministry of Defense. Armed Forces Covenant. 2017. Veteran: Key Facts URL: https://www.armedforcescovenant.gov.uk/wp-content/uploads/2016/02/Veterans-Key-Facts.pdf [accessed 2020-03-20]
2. Government of UK. London, UK: Ministry of Defence; 2019 Jan 10. Population Projections: UK Armed Forces Veterans Residing in Great Britain, 2016 to 2018 URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/775151/20190107_Enclosure_1_Population_Projections_-_UK_Armed_Forces_Veterans_residing_in_Great_Britain_-_2016_to_2028.pdf [accessed 2020-03-20]
3. Leightley D, Chui Z, Jones M, Landau S, McCrone P, Hayes RD, et al. Integrating electronic healthcare records of armed forces personnel: Developing a framework for evaluating health outcomes in England, Scotland and Wales. Int J Med Inform 2018 May;113:17-25 [FREE Full text] [doi: 10.1016/j.ijmedinf.2018.02.012] [Medline: 29602429]
4. Payne RA, Abel GA, Guthrie B, Mercer SW. The effect of physical multimorbidity, mental health conditions and socioeconomic deprivation on unplanned admissions to hospital: a retrospective cohort study. Can Med Assoc J 2013 Mar 19;185(5):E221-E228 [FREE Full text] [doi: 10.1503/cmaj.121349] [Medline: 23422444]
5. Simmonds S, Syddall H, Walsh B, Evandrou M, Dennison E, Cooper C, et al. Understanding NHS hospital admissions in England: linkage of Hospital Episode Statistics to the Hertfordshire Cohort Study. Age Ageing 2014 Sep;43(5):653-660 [FREE Full text] [doi: 10.1093/ageing/afu020] [Medline: 24598084]
6. Stevelink SA, Jones M, Hull L, Pernet D, MacCrimmon S, Goodwin L, et al. Mental health outcomes at the end of the British involvement in the Iraq and Afghanistan conflicts: a cohort study. Br J Psychiatry 2018 Dec;213(6):690-697 [FREE Full text] [doi: 10.1192/bjp.2018.175] [Medline: 30295216]

7. Fear NT, Jones M, Murphy D, Hull L, Iversen AC, Coker B, et al. What are the consequences of deployment to Iraq and Afghanistan on the mental health of the UK armed forces? A cohort study. Lancet 2010 May 22;375(9728):1783-1797. [doi: 10.1016/S0140-6736(10)60672-1] [Medline: 20471076]
8. Stevelink SA, Jones N, Jones M, Dyball D, Khera CK, Pernet D, et al. Do serving and ex-serving personnel of the UK armed forces seek help for perceived stress, emotional or mental health problems? Eur J Psychotraumatol 2019;10(1):1556552 [FREE Full text] [doi: 10.1080/20008198.2018.1556552] [Medline: 30693074]
9. Morgan VA, Jablensky AV. From inventory to benchmark: quality of psychiatric case registers in research. Br J Psychiatry 2010 Jul;197(1):8-10. [doi: 10.1192/bjp.bp.109.076588] [Medline: 20592426]
10. Mark KM, Leightley D, Pernet D, Murphy D, Stevelink SA, Fear NT. Identifying veterans using electronic health records in the United Kingdom: a feasibility study. Healthcare (Basel) 2019 Dec 19;8(1) [FREE Full text] [doi: 10.3390/healthcare8010001] [Medline: 31861575]
11. Leightley D, Williamson V, Darby J, Fear NT. Identifying probable post-traumatic stress disorder: applying supervised machine learning to data from a UK military cohort. J Ment Health 2019 Feb;28(1):34-41. [doi: 10.1080/09638237.2018.1521946] [Medline: 30445899]
12. Karstoft K, Statnikov A, Andersen SB, Madsen T, Galatzer-Levy IR. Early identification of posttraumatic stress following military deployment: Application of machine learning methods to a prospective study of Danish soldiers. J Affect Disord 2015 Sep 15;184:170-175. [doi: 10.1016/j.jad.2015.05.057] [Medline: 26093830]
13. Cambria E, White B. Jumping NLP curves: a review of Natural Language Processing Research [Review Article]. IEEE Comput Intell Mag 2014 May;9(2):48-57. [doi: 10.1109/mci.2014.2307227]
14. Fernandes AC, Dutta R, Velupillai S, Sanyal J, Stewart R, Chandran D. Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using Natural Language Processing. Sci Rep 2018 May 9;8(1):7426 [FREE Full text] [doi: 10.1038/s41598-018-25773-2] [Medline: 29743531]
15. Dalianis H. Clinical Text Mining: Secondary Use Of Electronic Patient Records. Cham: Springer; 2018.
16. Reeves RM, Ong FR, Matheny ME, Denny JC, Aronsky D, Gobbel GT, et al. Detecting temporal expressions in medical narratives. Int J Med Inform 2013 Feb;82(2):118-127. [doi: 10.1016/j.ijmedinf.2012.04.006] [Medline: 22595284]
17. Gundlapalli AV, Carter ME, Palmer M, Ginter T, Redd A, Pickard S, et al. Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans. AMIA Annu Symp Proc 2013;2013:537-546 [FREE Full text] [Medline: 24551356]
18. Mowery DL, Chapman BE, Conway M, South BR, Madden E, Keyhani S, et al. Extracting a stroke phenotype risk factor from Veteran Health Administration clinical reports: an information content analysis. J Biomed Semantics 2016;7:26 [FREE Full text] [doi: 10.1186/s13326-016-0065-1] [Medline: 27175226]
19. Gundlapalli AV, Jones AL, Redd A, Divita G, Brignone E, Pettey WB, et al. Combining Natural Language Processing of electronic medical notes with administrative data to determine racial/ethnic differences in the disclosure and documentation of military sexual trauma in veterans. Med Care 2019 Jun;57(Suppl 6 Suppl 2):S149-S156. [doi: 10.1097/MLR.0000000000001031] [Medline: 31095054]
20. Perera G, Broadbent M, Callard F, Chang C, Downs J, Dutta R, et al. Cohort profile of the South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. BMJ Open 2016 Mar 1;6(3):e008721 [FREE Full text] [doi: 10.1136/bmjopen-2015-008721] [Medline: 26932138]
21. Downs JM, Ford T, Stewart R, Epstein S, Shetty H, Little R, et al. An approach to linking education, social care and electronic health records for children and young people in South London: a linkage study of child and adolescent mental health service data. BMJ Open 2019 Jan 29;9(1):e024355 [FREE Full text] [doi: 10.1136/bmjopen-2018-024355] [Medline: 30700480]
22. Velupillai S, Hadlaczky G, Baca-Garcia E, Gorrell GM, Werbeloff N, Nguyen D, et al. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front Psychiatry 2019;10:36 [FREE Full text] [doi: 10.3389/fpsyt.2019.00036] [Medline: 30814958]
23. Jackson RG, Patel R, Jayatilleke N, Kolliakou A, Ball M, Gorrell G, et al. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ Open 2017 Jan 17;7(1):e012012 [FREE Full text] [doi: 10.1136/bmjopen-2016-012012] [Medline: 28096249]
24. Kovalchuk Y, Stewart R, Broadbent M, Hubbard TJ, Dobson RJ. Analysis of diagnoses extracted from electronic health records in a large mental health case register. PLoS One 2017;12(2):e0171526 [FREE Full text] [doi: 10.1371/journal.pone.0171526] [Medline: 28207753]
25. Mueller C, Perera G, Hayes R, Shetty H, Stewart R. Associations of acetylcholinesterase inhibitor treatment with reduced mortality in Alzheimer's disease: a retrospective survival analysis. Age Ageing 2018 Jan 1;47(1):88-94. [doi: 10.1093/ageing/afx098] [Medline: 28655175]
26. NIHR Maudsley Biomedical Research Centre (BRC). Clinical Record Interactive Search (CRIS) URL: https://www.maudsleybrc.nihr.ac.uk/facilities/clinical-record-interactive-search-cris/ [accessed 2020-03-20]

27. Juckett D. A method for determining the number of documents needed for a gold standard corpus. J Biomed Inform 2012 Jun;45(3):460-470 [FREE Full text] [doi: 10.1016/j.jbi.2011.12.010] [Medline: 22245601]
28. Leng CJ, South B, Shen S. Orbit. Utah: University of Utah and SLC VA; 2011. eHOST: The Extensible Human Oracle Suite of Tools URL: https://orbit.nlm.nih.gov/browse-repository/software/nlp-information-extraction/62-ehost-the-extensible-human-oracle-suite-of-tools [accessed 2020-03-24]
29. Loper E, Bird S. NLTK: the Natural Language Toolkit. In: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics - Volume 1. USA: Association for Computational Linguistics; 2002 Presented at: ETMTNLP'02; July 2002; Morristown p. 63-70. [doi: 10.3115/1118108.1118117]
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825-2830 [FREE Full text]
31. Leightley D, Darby J, Baihua L, McPhee J, Yap MM. Human Activity Recognition for Physical Rehabilitation. In: Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics. 2013 Presented at: SMC'13; October 13-16, 2013; Manchester, UK. [doi: 10.1109/SMC.2013.51]
32. Leightley D, McPhee JS, Yap MH. Automated analysis and quantification of human mobility using a depth sensor. IEEE J Biomed Health Inform 2017 Jul;21(4):939-948. [doi: 10.1109/JBHI.2016.2558540] [Medline: 27254874]
33. Ahad M, Tan J, Kim H, Ishikawa S. Human Activity Recognition: Various Paradigms. In: Proceedings of the 2008 International Conference on Control, Automation and Systems. COEX, Seoul, Korea: IEEE; 2008 Presented at: ICCAS'08; October 14-17, 2008; Seoul, Korea p. 1896-1901. [doi: 10.1109/ICCAS.2008.4694407]
34. Cunningham R, Sánchez M, May G, Loram I. Estimating full regional skeletal muscle fibre orientation from B-mode ultrasound images using convolutional, residual, and deconvolutional neural networks. J. Imaging 2018 Jan 29;4(2):29. [doi: 10.3390/jimaging4020029]
35. Leightley D. GitHub. Military Service Identification Tool URL: https://github.com/DrDanL/kcmhr-msit [accessed 2020-03-24]
36. Pineda AL, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui FR. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform 2015 Dec;58:60-69 [FREE Full text] [doi: 10.1016/j.jbi.2015.08.019] [Medline: 26385375]
37. Cronin RM, Fabbri D, Denny JC, Rosenbloom ST, Jackson GP. A comparison of rule-based and machine learning approaches for classifying patient portal messages. Int J Med Inform 2017 Sep;105:110-120 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.06.004] [Medline: 28750904]
38. Stewart R. The big case register. Acta Psychiatr Scand 2014 Aug;130(2):83-86. [doi: 10.1111/acps.12279] [Medline: 24730985]
39. Stewart R, Soremekun M, Perera G, Broadbent M, Callard F, Denis M, et al. The South London and Maudsley NHS Foundation Trust Biomedical Research Centre (SLAM BRC) case register: development and descriptive data. BMC Psychiatry 2009 Aug 12;9:51 [FREE Full text] [doi: 10.1186/1471-244X-9-51] [Medline: 19674459]
40. Wickersham A, Petrides PM, Williamson V, Leightley D. Efficacy of mobile application interventions for the treatment of post-traumatic stress disorder: A systematic review. Digit Health 2019;5:2055207619842986 [FREE Full text] [doi: 10.1177/2055207619842986] [Medline: 31019722]

Abbreviations

CRIS: Clinical Record Interactive Search
EHR: electronic health care record
MSIT: Military Service Identification Tool
NHS: National Health Service
NIHR: National Institute of Health Research
NLP: natural language processing
SLaM: South London and Maudsley
SQL: structured query language
tf-idf: term frequency-inverse document frequency

Edited by G Eysenbach; submitted 13.08.19; peer-reviewed by S Purkayastha, G Molina Recio, S Butler, K Goniewicz; comments to author 03.10.19; revised version received 11.12.19; accepted 26.01.20; published 25.05.20.

Please cite as:
Leightley D, Pernet D, Velupillai S, Stewart RJ, Mark KM, Opie E, Murphy D, Fear NT, Stevelink SAM
The Development of the Military Service Identification Tool: Identifying Military Veterans in a Clinical Research Database Using Natural Language Processing and Machine Learning
JMIR Med Inform 2020;8(5):e15852
URL: http://medinform.jmir.org/2020/5/e15852/
doi:10.2196/15852
PMID:32348287

Original Paper

Determining the Topic Evolution and Sentiment Polarity for Albinism in a Chinese Online Health Community: Machine Learning and Social Network Analysis

Qiqing Bi[1,2,3*], MS; Lining Shen[1,2,3*], PhD; Richard Evans[4], PhD; Zhiguo Zhang[1,2], PhD; Shimin Wang[1,2], MS; Wei Dai[1], MS; Cui Liu[1], MS

[1]School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science & Technology, Wuhan, China

[2]Hubei Provincial Research Center for Health Technology Assessment, Wuhan, China

[3]Institute of Smart Health, Huazhong University of Science & Technology, Wuhan, China

[4]College of Engineering, Design and Physical Sciences, Brunel University London, London, United Kingdom

[*]these authors contributed equally

Corresponding Author:
Lining Shen, PhD
School of Medicine and Health Management
Tongji Medical College
Huazhong University of Science & Technology
No 13 Hangkong Road
Wuhan, 430030
China
Phone: 86 027 83692730
Email: sln2008@hust.edu.cn

Abstract

Background: There are more than 6000 rare diseases in existence today, with the number of patients with these conditions rapidly increasing. Most research to date has focused on the diagnosis, treatment, and development of orphan drugs, while few studies have examined the topics and emotions expressed by patients living with rare diseases on social media platforms, especially in online health communities (OHCs).

Objective: This study aimed to determine the topic categorizations and sentiment polarity for albinism in a Chinese OHC, Baidu Tieba, using multiple methods. The OHC was deeply mined using topic mining, social network analysis, and sentiment polarity analysis. Through these methods, we determined the current situation of community construction, identifying the ongoing needs and problems experienced by people with albinism in their daily lives.

Methods: We used the albinism community on the Baidu Tieba platform as the data source in this study. Term frequency–inverse document frequency, latent dirichlet allocation models, and naive Bayes were employed to mine the various topic categories. Social network analysis, which was completed using the Gephi tool, was employed to analyze the evolution of the albinism community. Sentiment polarity analysis was performed using a long short-term memory algorithm.

Results: We identified 8 main topics discussed in the community: daily sharing, family, interpersonal communication, social life and security, medical care, occupation and education, beauty, and self-care. Among these topics, daily sharing represented the largest proportion of the discussions. From 2012 to 2019, the average degree and clustering coefficient of the albinism community continued to decline, while the network center transferred from core communities to core users. A total of 68.43% of the corpus was emotional, with 35.88% being positive and 32.55% negative. There were statistically significant differences in the distribution of sentiment polarity between topics (P<.001). Negative emotions were twice as high as positive emotions in the social life and security topic.

Conclusions: The study reveals insights into the emotions expressed by people with albinism in the Chinese OHC, Baidu Tieba, providing health care practitioners with greater appreciation of the current emotional support needed by patients and the patient experience. Current OHCs do not exert enough influence due to limited effective organization and development. Health care sectors should take greater advantage of OHCs to support vulnerable patients with rare diseases to meet their evidence-based needs.

(JMIR Med Inform 2020;8(5):e17813) doi:10.2196/17813

KEYWORDS

albinism; rare diseases; topic mining; social network analysis; sentiment polarity; online health community; machine learning

Introduction

Background

Rare diseases are considered conditions that affect a limited amount of people, typically less than 1 in 2000 individuals. Albinism is a type of rare disease related to a variable hypopigmentation phenotype, where patients experience partial or complete absence of pigment in their skin, eyes, and hair [1]. Despite advances in genomic technology and medicines, many individuals affected with rare diseases remain undiagnosed, and some never receive a definitive diagnosis [2]. A diagnosis with a rare disease is extremely likely to cause economic, psychosocial, and physical burden on the patient and family members [3]. Research demonstrates that parents of children with rare genetic disorders present feelings of social isolation, anxiety, fear, anger, and uncertainty [4] and experience high levels of physical and emotional strain [5].

Related Research

Over the last decade, rare disease research has received considerable attention in health care studies, with exploration typically focusing on 1 of 3 main areas: etiology, diagnosis, and treatment [6]. In recent years, rare disease research has also straddled other disciplines, including policy improvement, sociology, psychology, and ethics. For example, Abbas et al [7] reported that the European Union and United States have adopted policies and regulations aimed at improving orphan drug availability over the past 20 years, but that only 16 countries had an orphan drug or rare disease plan in place. Rodwell and Ayme [8] reviewed the political frameworks of European countries to demonstrate how legislation has created a dynamic that is progressively improving health care for patients with rare diseases. Dharssi et al [9] found that patient communities are being used to promote and drive the establishment and adoption of legislation and programs to improve rare disease care. Gomes [10] discussed the construction of social identity, mutual recognition, and the specific demands for recognition of people with rare conditions from 3 sociological perspectives.

Online Health Communities

Online health communities (OHCs) have become a popular means for individuals to obtain support and connect with others online when experiencing illness, especially patients with similar diagnoses [11]. An increasing amount of literature related to OHCs documents widespread concerns from scholars worldwide. Some researchers have focused mostly on social networks and user behaviors. For example, Huh et al [12] conducted open coding analysis using interview data and cluster analysis to determine that 4 types of persona exist in OHCs: caretakers, opportunists, scientists, and adventurers. Lu el al [13] investigated health care social media use from different stakeholder perspectives using content analysis. Others have concentrated on knowledge sharing and value creation. For example, Yan et al [14] proposed a benefit versus cost knowledge sharing model for OHCs. Guo et al [15] conducted an empirical investigation into the relationship between professional capital and exchange returns in OHCs. In addition, health interventions have been reported based on OHCs. Naslund et al [16] established that people with serious mental health illnesses reported benefits from interacting with peers online, experiencing greater social connectedness. Most existing OHC research has examined chronic diseases, such as cancer, diabetes, AIDS, and severe mental disorders, using large patient populations and relating more to social concerns [17-20]. Furthermore, social media tools have been studied, such as Wechat Official Accounts [21] and SentiHealth-Cancer [22]. However, there are few studies that have focused on OHCs for rare diseases. Davies et al [23] found that online surveys for stakeholder groups may provide new insights into rare conditions and their management relatively quickly, with the possibility of rapid translation into health care intervention management and policy development. Although the number of patients with rare diseases is limited, some scholars have pointed out that patients with such conditions require increased social support networks [24].

Objectives

The main type of albinism is oculocutaneous albinism, which is a group of conditions that affect the coloring (pigmentation) of the skin, hair, and eyes. Long-term exposure to the sun can greatly increase the risk of skin damage and cancer [25]. Melanin deficiency causes a series of abnormalities in the eyes, such as severe low vision, photophobia, and nystagmus. Due to its special phenotype, the psychological development of patients with albinism is affected [26]. The worldwide prevalence of oculocutaneous albinism is estimated to be 1 in 17,000 [27]. In the Chinese Han ethnic group population of the Shandong province in China, the prevalence is approximately 1 in 18,000, or roughly 3.80% of the population [28]. In addition to the general characteristics of more typical rare diseases, albinism has a certain uniqueness and patient base. Current academic research into albinism has focused on etiology [29], pathology [30,31], diagnosis [32-35], sociology [36,37], and albinism in animals [38,39].

To our knowledge, no studies exist on albinism-based OHCs, aimed at deeply detecting the prevailing topics, their change over time, and sentiment polarity (ie, sentimental expressions of albinism patients and the distribution of different sentiments). This study aimed to guide the academic community to focus more on rare diseases in albinism OHCs. Specifically, this study aimed to answer 3 research questions. What is the topic evolution for albinism in OHCs? What are the characteristics of albinism social networks in OHCs? What is the sentiment polarity of albinism in OHCs?

Methods

Sample and Data Collection

Few OHCs for albinism exist in China, with most related to social media, such as Tencent QQ, WeChat, and Baidu Tieba [40]. Baidu Tieba is the largest Chinese communication platform for discussion and the posting of questions [41], with data being readily available and considered high quality. This platform contains millions of online communities targeted at specific topics. The Baidu albinism community has over 300,000 registered users. Accordingly, we designed a web spider using Python 3.7 [42] Scrapy [43] to crawl the records dated from January 30, 2007 to March 14, 2019, including a total of 5802 posts, 45,181 comments, and 3977 active users. The dataset contains content of posts and the complete text of comments, as shown in Textbox 1. Given that some data collected before 2012 were severely lost and fragmented, the dataset from 2012 to 2019 was eventually selected for subsequent analysis. In addition, the following user-posted content was also removed: non-text content (eg, video, music, picture) or content with missing author and time fields. The final dataset included 5110 posts, 35,414 comments, and 3188 active users. The process for identifying data for subsequent analysis is shown in Figure 1. Moreover, we categorized users who had not used the albinism community for more than 1 year as "lost users," and users who had used the community more regularly as "new users."

Textbox 1. Data fields extracted from the online albinism community.

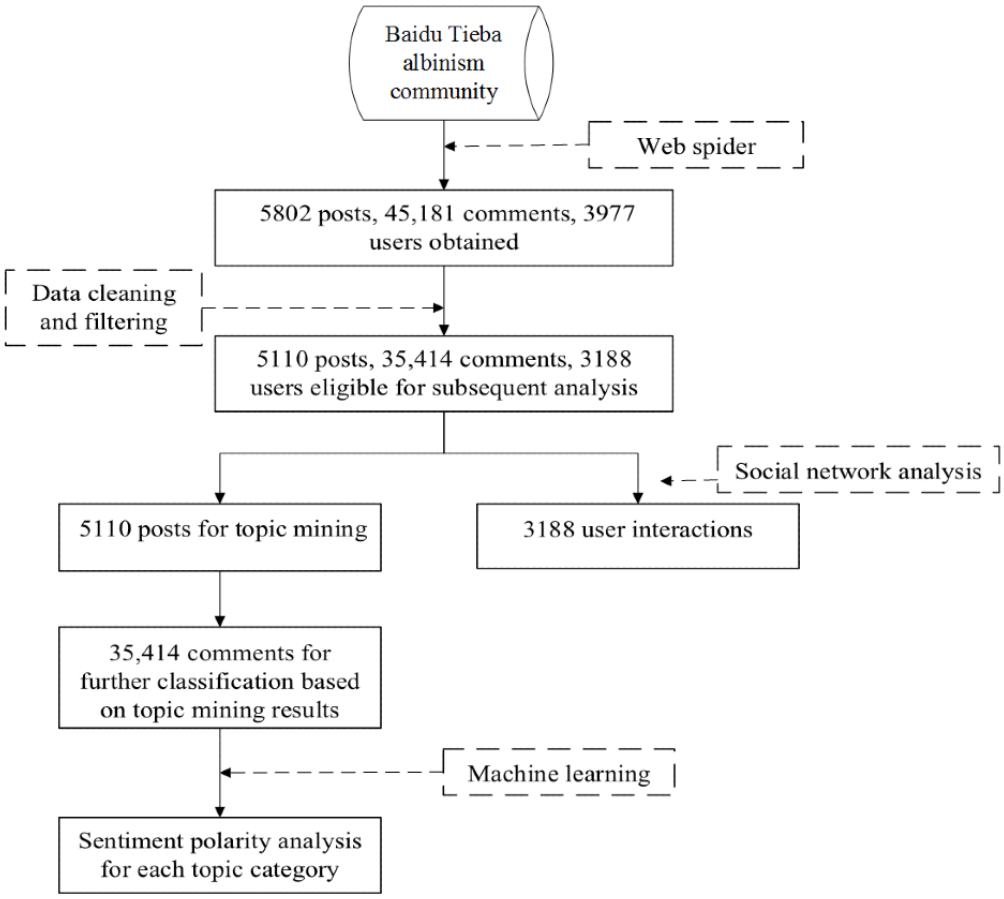
Albinism_Post

- Post_id (post id)
- Post_title (post title)
- Author_id (author's id)
- Content (post content)
- Time (post time)
- Reply_num (number of replies)
- URL (URL of the post)

Albinism_Comment

- Comment_id (comment id)
- Post_id (post to which the comment belongs)
- Author_id (author's id)
- Content (comment content)
- Time (comment time)
- Floor (the floor in its post, which represents a comment from a user, and the floor number is order of user comments)

Figure 1. Flowchart for identifying data from the online albinism Baidu community for subsequent analysis.



Data Analysis Methods

Topic Mining

To ensure the amount and accuracy of topic mining, this study used the title and comments as the topic mining corpus. After data cleansing, the dataset for topic mining contained 10,220 corpora. First, Jieba 0.39 [44] in Python 3.7, the Chinese word segmentation tool, was employed for word segmentation. Owing to the particularity of albinism in the medical field, we used the International Statistical Classification of Diseases and Related Health Problems, 10th Revision and Chinese Medical Subject Headings to expand the lexical dictionary for intervention. In addition, based on the stop word list of the Harbin Institute of Technology in China, our stop word list was continuously updated through the results throughout the experiment.

Then, we combined term frequency–inverse document frequency and latent dirichlet allocation (LDA) [45] for topic mining; the number of topics was identified based on the perplexity [46]. Here, LDA, the most common method for topic modelling, is a generalization of probabilistic latent semantic indexing [47]. Perplexity is a common criterion for evaluating the effectiveness of language models [48]. Due to each topic in the LDA results containing multiple types of topic information, two research assistants (RAs) with medical backgrounds were hired to independently annotate each LDA category with 1-3 labels. Then, the RAs evaluated the results independently to reach consensus, with discussions for any discrepancies or disagreements joined by the first author of this study. Subsequently, the assigned labels were combined, deduplicated, and reclassified to form the final classification label. Moreover, a naive Bayes (NB) model was used, which performs well with small-scale data and can handle multiple classification tasks commonly used for text classification [49]. Therefore, on the basis of the new classification label, a NB classifier was created to classify all posts, with a precision rate of 0.889, recall rate of 0.915, and F1 score of 0.902. Finally, each comment was merged into the topic of the corresponding post; the topic classification for the full corpus was implemented since the comment text was short and the topic information was limited.

Social Network Analysis

A social network is the integration of social relationships. With the increase in popularity of social media sites, scholars and practitioners have aimed to understand the behaviors of people using such platforms [50,51]. Gephi, a social network visualization software, is used in various disciplines. One of its key features is the ability to display the spatialization process [52]. Gephi 0.9.2 [53] was employed in this study to analyze the topology of the interaction between 3188 users, based on the community mining algorithm built in the software [54], which can detect the potential community of users. As the results of the analysis for all user data were ambiguous, we identified a 2-year interval to explore the dynamic evolution of the community structure to better reflect the users' activity. To better reflect the social network characteristics of the albinism bar, we compared it to the random networks with the same number of nodes based on several basic indicators, including average degree, network diameter, number of communities, clustering coefficient, and average path length. The average

degree represents the average distance between nodes. The clustering coefficient is a coefficient indicating the degree of node aggregation in a graph. The average path length is the average shortest distance between all pairs of nodes in the network.

Sentiment Polarity Analysis

Sentiment polarity analysis, commonly used in academia, mainly includes a sentiment dictionary and machine learning. And the frontier branch of machine learning is deep learning [55,56]. At present, the enhanced version of machine learning algorithms is widely used in sentiment analysis [57,58]. Therefore, we selected 4 representative training classifiers of machine learning algorithms, including NB, support vector machine, convolutional neural network, and long short-term memory. Sentiment polarity was divided into 3 polarities: positive, neutral, and negative. We first randomly chose more than 4000 corpora and then marked them with one of these 3 sentiment polarities using Colabeler (Hangzhou Kuaiyi Technology Co Ltd, Hangzhou, Zhejiang, China), a labeling program. Then, we selected 1000 records marked with one sentiment polarity from 4000 corpora for the sentiment classification model training. The corpus that stated objective facts was marked as neutral. The others that contained obvious sentiment words and emotions were marked as positive or negative. In this process, we referred to the Hownet sentiment lexicon [59] from the China National Knowledge Infrastructure and the Chinese sentiment lexicon and sentiment analyzer from the National Taiwan University School of Dentistry [60]. As shown in Table 1, the long short-term memory classifier performed best in the testing of sentiment polarity for the remaining corpora, in comparison with the 3 alternative machine learning algorithms. Finally, the differences in sentiment distribution between topics was verified using a Chi-square test executed in SPSS 20.0 (IBM Corp, Armonk, NY).

Table 1. Performance of the models for sentiment polarity classification.

	Precision	Recall	F1 score
NB[a]	0.798	0.835	0.816
SVM[b]	0.853	0.822	0.837
CNN[c]	0.801	0.823	0.812
LSTM[d]	0.916	0.916	0.916

[a]NB: naive Bayes.

[b]SVM: support vector machine.

[c]CNN: convolutional neural network.

[d]LSTM: long short-term memory.

Results

Basic Statistical Information

From 2012 to 2019, the number of posts and comments showed the same trend: they increased during the early years of the study, reached a peak in 2015, and subsequently declined (Figure 2). The findings revealed that the users preferred to use the albinism community after 6:00 pm, with all other times similar in frequency of use; there were only two small peaks at lunch and dinner times, as shown in Figure 3.

Figure 2. Posts and comments about albinism in the online community in 2012-2019.

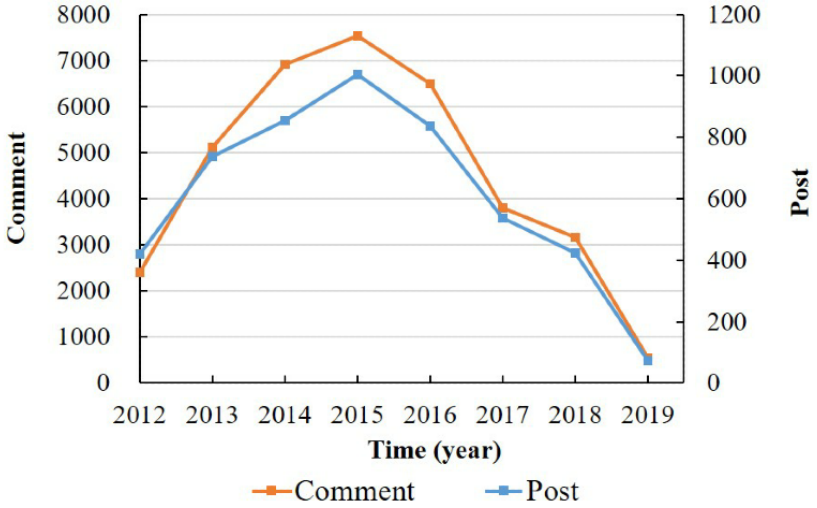


Figure 3. Distribution of the comments in the online albinism community by hour of the day.

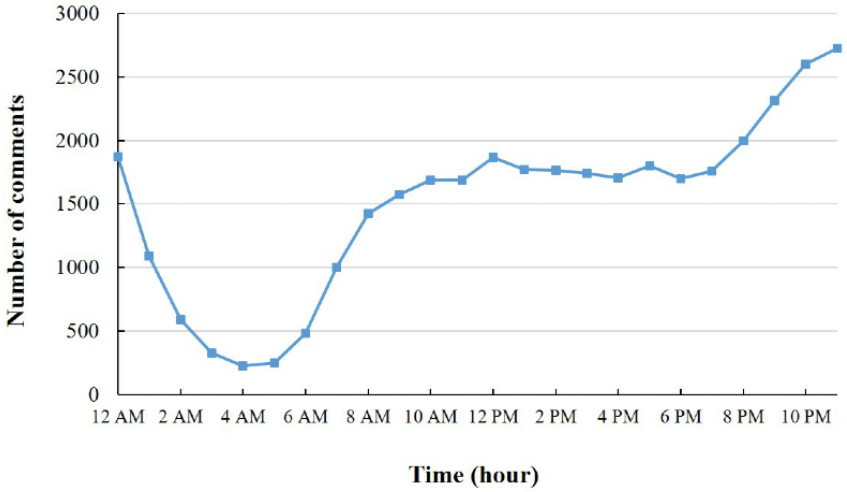


Figure 4 shows that the number of active users increased during the early years of the study period but peaked in 2016 and then declined. Furthermore, the number of "lost users" increased each year, indicating that the speed of user abandonment increased, whereas the number of "new users" increased at the beginning and then decreased at a faster rate than it increased. The superposition of the two curves shows a significant decline in the number of active community members. The trend remained obvious even after omitting the 2019 data. Figure 5 presents the average number of posts submitted by users each year, showing a decreasing trend year by year.

Figure 4. Number of users in the online albinism community per year.

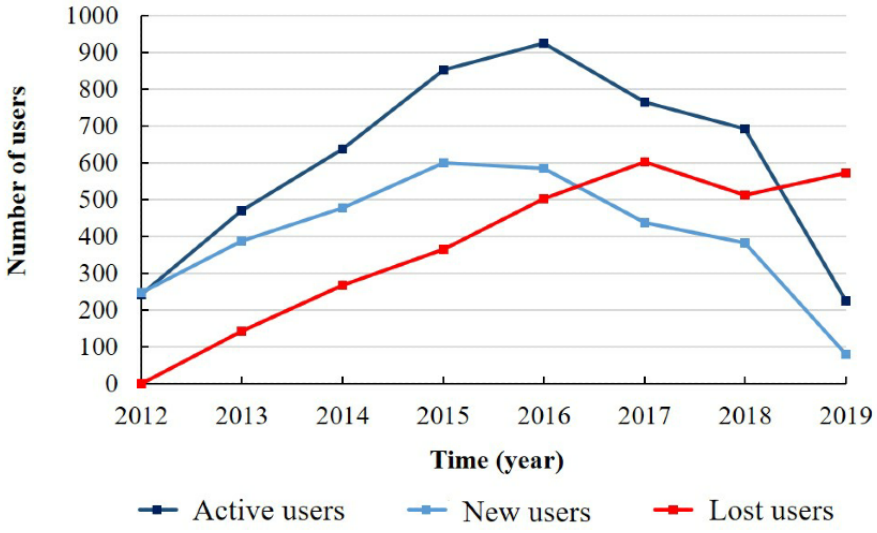
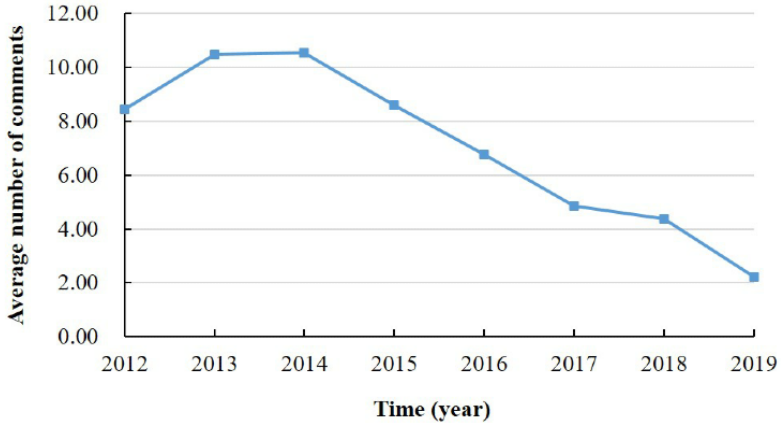


Figure 5. Average number of comments posted in the albinism community each year.



Topic Evolution

As shown in Figure 6, the lowest perplexity was 36, which determined the value of the parameter num_topics of the LDA document topic generation model. For the details of these 36 categories, see Multimedia Appendix 1. Moreover, after merging and sorting, the final classification labels were formed, with a total of 8 categories, shown in Table 2.

Figure 6. Latent dirichlet allocation model topic number in a perplexity diagram.

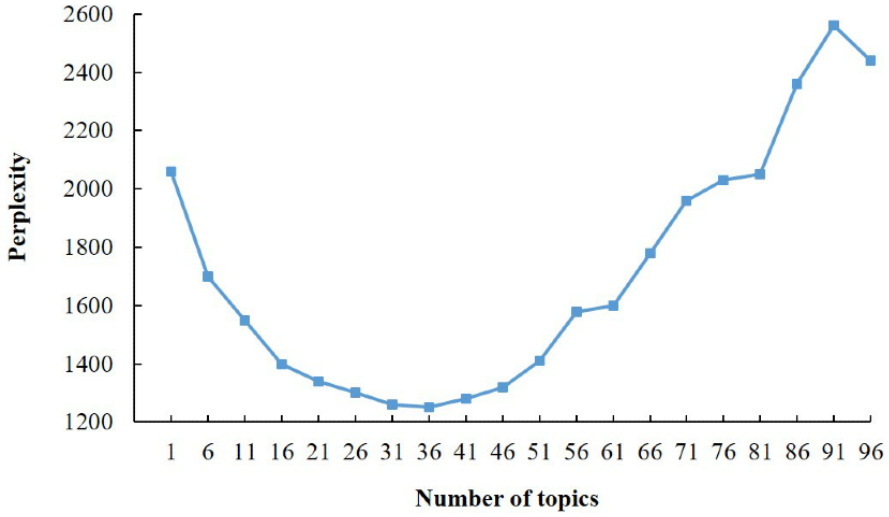


Table 2. The resulting 8 categories for the posts about albinism in the online community.

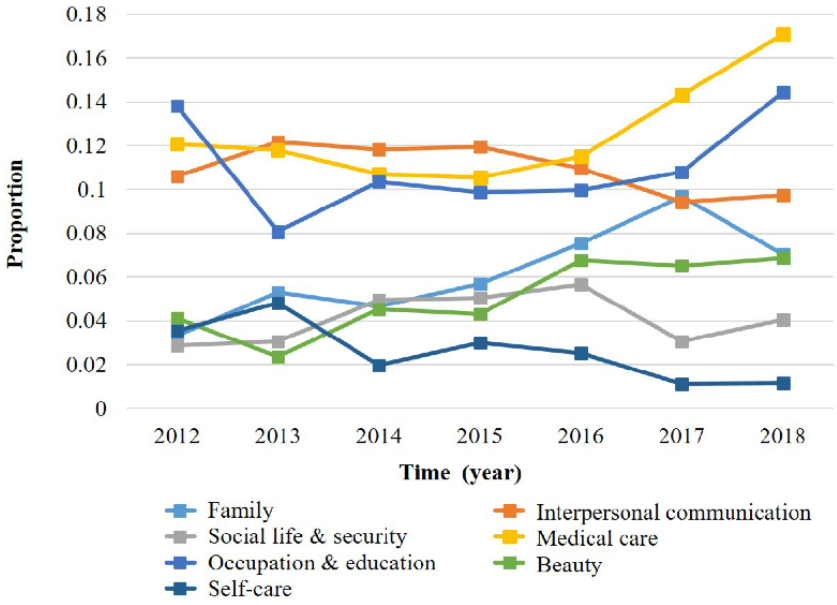
Number	Category name	Description	Examples
1	Daily sharing	Sharing of daily life experiences (not included in topics 2-8)	The weather is really good today! It's unlucky to lose money.
2	Family	Sharing of daily life experiences from the perspective of family members of people with albinism	I have an angel baby. My child is diagnosed with albinism, so desperate.
3	Interpersonal communication	Social contact requests	Let's make friends! Are there friends from Beijing? This is my QQ number.
4	Social life & security	Discussion of social impact or social commonality	How do I apply for a disability certificate? Where can I get free vision glasses?
5	Medical care	Medical issues, such as treatment, examination, and protection	What medical examination is needed? What about nystagmus?
6	Occupation & education	Issues related to occupation or education	How about the income of the massage industry? Does albinism not affect school?
7	Beauty	Issues related to hair care, dyeing of hair, or makeup	Can people with albinism dye their hair? The younger sister's makeup is really beautiful.
8	Self-care	Other issues related to daily life (not included in topics 3-7)	How to repair the computer? How to register a game account?

After all the comments were classified as topics according to the results of the topic category of the posts, the daily sharing category accounted for the largest proportion (17,010/35,414, 48.03%) of the total comments, indicating that users were open to expressing their feelings and daily life through social media. Medical care was the second most common subject discussed by users, accounting for 12.04% (4264/35,414) of the total comments posted. With regards to this category, genetic testing, prenatal testing, vision protection, skin protection, and treatment were the major topics discussed. An indepth analysis of the corpus found that users were confused about methods of protection and how to obtain them. Interpersonal communication was the third most discussed topic among users, accounting for 11.20% (3966/35,414) of the comments posted. This reflects the social attributes of Baidu Tieba, with users searching for suitable companions based on region, age, hobby, and disease severity. There were also numerous exchanges in the occupation & education category, representing 10.53% (3729/35,414) of the total comments; these two aspects were observed to be a severe annoyance for people with albinism. Visual impairment

and fragile skin interfere with occupation and education. The family and beauty categories accounted for 6.17% (2185/35,414) and 5.00% (1771/35,414), respectively, of the posted comments. The family category reflected the emotional expression among family members. As the issues for family members are also involved in the medical care and social life & security aspects for people with albinism, the proportion here is slightly lower. Beauty reflected the patient's pursuit of appearance and positive attitude towards life, which can alleviate some practical issues. The categories with the lowest number of comments were social life & security (1558/35,414, 4.40%) and self-care (931/35,414, 2.63%). The social life & security category included public welfare activities, public events, policies, and regulations, representing the maintenance of patients' rights and interests.

The absolute number of each topic corpus was affected by the overall trend. Figure 7 shows the change in the proportion of 7 topic categories from 2012 to 2018; the daily sharing category was excluded because its proportion far exceeded those of the other categories. It can be intuitively seen that the number of posts within the medical care, occupation & education, and beauty categories dynamically increased during the study period. Among the categories, the increase in the number of posts in the medical care category is the most obvious. These 3 categories represent a certain degree of disease experience sharing, indicating that the online albinism community provided an effective platform for patients to solve problems to some extent. The number of posts in the family category also experienced an upward trend but declined in 2018. The number of posts in the other 3 categories fluctuated or declined to varying degrees during the study period.

Figure 7. Topic evolution by year, with each category reported as a proportion of the total comments per year.



Social Network Structure

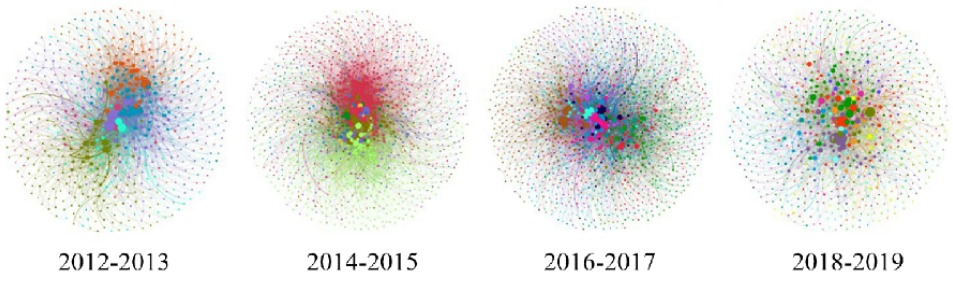
As shown in Table 3, we observed that the average degree and clustering coefficient continued to decrease, while the network diameter, number of communities, and average path length increased. However, these results are better than that of random networks with the same size from the perspective of user interaction. This shows that there is a small world effect between users, which can form effective communication, but this effect is gradually decreasing.

Table 3. Basic statistics for the social network analysis, compared with those of a random network.

Year	Number of users	Average degree		Network diameter		Number of communities		Clustering coefficient		Average path length	
		Study network	Random network	Study network	Random network	Study network	Random network	Study network	Random network	Study network	Random network
2013	629	5.67	16.08	7	10	6	8	0.210	0.026	3.20	2.58
2014	951	7.00	23.70	8	10	9	10	0.176	0.025	3.21	2.45
2015	1268	6.78	31.73	9	10	13	11	0.136	0.025	3.36	2.36
2016	1472	5.99	36.98	8	9	13	10	0.113	0.025	3.51	2.31
2017	1415	4.81	35.46	10	10	13	10	0.097	0.025	3.81	2.32
2018	1212	3.98	30.37	11	9	16	10	0.077	0.025	4.35	2.37
2019	796	3.29	19.79	14	11	23	10	0.072	0.025	4.59	2.49

Figure 8 presents the evolution of the community structure from 2012 to 2019, which reflects the distribution characteristics of core edge. The node represents the users, and the node size is proportional to the degree. Different communities are distinguished by color. The edge represents the comment relationship between users. The structural changes occurred from the core community to the core user as the principal part in evidence. From 2012 to 2016, the number of communities increased in the central region. Meanwhile, the scale expanded, and the structure matured. From 2016 to 2019, the community replaced by core users has become blurred in the central region, while the number of core users has increased significantly.

Figure 8. Changes in the community structure over time.



Distribution of Sentiment Polarity

Daily sharing was the most active category (12,581/17,010, 73.96%) for expressing emotions, with positive emotions being observed the most often (7170/17,010, 42.15%), as shown in Table 4. When users encounter events that affect their emotions in their daily lives, they tend to vent through social media. The online albinism community is seen to provide a platform for confiding with other people with albinism and their families. In addition, the medical care category had the highest proportion (1671/4264, 39.19%) of negative emotions. Most people with albinism have skin and vision dysfunction, which causes a number of practical issues that affect quality of life. The negative emotions expressed in the medical care category arose from issues mainly related to anxiety and worry, such as “Does this disease only affect white-skinned people?” and “How do I deal with blurred vision?” With regards to the family category, there were many similar statements such as “I cry at home every day” or “I don’t know what to do” that conveyed feelings of sadness, confusion, and helplessness. Moreover, the social life & security category had a high proportion of negative emotions (588/1558, 37.74%), twice that of the number of positive emotions. This category is concerned mostly with public benefits such as the distribution of visual aids, health education, and offline activities. However, many posts referred to the handling and grading of disability certificates, social discrimination issues, and medical insurance, all of which are likely to increase negative emotions. In addition, the statistical test results showed a statistically significant difference in the distribution of sentiment polarity between topic categories (χ^2_{14}=1083.368, P<.001).

Table 4. Results of the sentiment polarity analysis results for the 8 topic categories.

Topic category	Positive, n (%)	Neutral, n (%)	Negative, n (%)
Daily sharing	7170 (42.15)	4429 (26.04)	5411 (31.81)
Family	609 (27.87)	888 (40.64)	688 (31.49)
Interpersonal communication	1321 (33.30)	1660 (41.86)	985 (24.84)
Social life & security	286 (18.36)	684 (43.90)	588 (37.74)
Medical care	1327 (31.12)	1266 (29.69)	1671 (39.19)
Occupation & education	1125 (30.17)	1313 (35.21)	1291 (34.62)
Beauty	617 (34.84)	551 (31.11)	603 (34.05)
Self-care	251 (26.96)	390 (41.89)	290 (31.15)

The number of posts with negative emotions in the family, occupation & education, and self-care categories was slightly higher than the number of posts with positive emotions. Therefore, we can infer that users encounter obstacles in family life, employment, and education. The interpersonal communication category had more posts with positive emotions (1321/3966, 33.30%) than with negative emotions (983/3966, 24.84%). Meeting acquaintances is one of the main reasons that people with albinism join OHCs. Finally, there was no significant difference in the proportion of posts with positive (617/1771, 34.84%) or negative (603/1563, 34.05%) emotions in the beauty category, indicating that the user’s mood was relatively stable when talking about makeup or hair coloring, for example.

Discussion

Principal Findings

This study explored the topic characteristics and sentiment distribution for an albinism community in the Baidu Tieba OHC from multiple dimensions using LDA, social network analysis, and sentiment polarity analysis. There were 8 hot topics in the

communication within the community, of which the daily sharing topic category represented the largest proportion. The social network structure was not stable. The importance of core users was gradually emerging. Emotional differences were demonstrated in distinct topics, implying varying user attitudes and statuses.

Solve Practical Problems

First, our study demonstrated that users desire to solve practical problems using OHCs. As observed, patients are used to asking for help from people with similar experiences. The increasing proportion of topics on medical care, occupation & education, and beauty was obvious. Among these topic categories, medical care, including prenatal care and diagnosis, was the category that the most users were concerned with, and patients with albinism did not know where to go and what to do, causing anxiety and stress. This suggests that patients would appreciate more professional support, even a cure. In addition, physical defects and social discrimination seriously affected the quality of life of patients with albinism. They continue to demand ways to ease, as much as possible, their daily lives, protecting their rights and interests. Furthermore, users want to relieve social issues by using OHCs to meet people in similar situations. Surprisingly, we found that offline gatherings were mentioned in the original corpora, which is also helpful for further communication between patients. Our results also show that there are relatively close communities of users, which are conducive to the transmission and resolution of information, and the role of core users is gradually increasing across boundaries of smaller communities.

Another survey reported that 62% of respondents recognized the diagnosis, and 69% discussed online information with their physician [61]. Obviously, the use of the internet for health care interactions may represent a necessity for patients with rare diseases to better manage their complex health needs [62]. Furthermore, the creation of online communities for patients and caregivers who share information about their disease may empower them and facilitate participation in clinical trials [63,64]. However, albinism communities do not clearly identify doctors from whom users can seek professional help.

Improve User Participation and Loyalty

Second, measures should be taken to improve user participation and loyalty in OHCs for albinism. Actual participation in albinism communities is <2% (3977/300,000), which is far less than the number of identified albinism patients. Most users belong to the diving type, indicating that the content in the community does not attract them or they do not have the courage to express opinions in the current environment. Our results show a serious loss of users that has been sustained throughout the past few years. The average number of annual comments continues to decline, and users' expectations and interest in participating with such communication decrease. It should be noted that this community is likely to disappear in the future, if nothing is done to improve participation. Credibility is a matter of great concern. As commonly agreed, the accuracy and perceived credibility of OHCs is pivotal in facilitating social relationships [65]. A positive correlation also exists between community communication activity and information quality [66]. Therefore, low user participation and loyalty reflect this crisis in the albinism community. The results of the social network analysis show that the influence of core users is gradually expanding, which provides opportunities for professionals to influence the public. However, due to the decline in the overall influence, it is difficult for us to clearly understand the albinism community within this context, especially in the communication environment led by medical staff and specialists.

Express Feelings

Third, patients with albinism are inclined to express their feelings, especially negative feelings, in OHCs. The combination of topic mining and sentiment polarity analysis revealed the concerns of users and their attitudes towards various issues. The sentiment analysis of the whole corpus showed that 68.42% of posts were emotional; there were 5 topics for which a negative sentiment was more prevalent than a positive sentiment. Therefore, users are used to expressing their feelings through the internet. OHCs provide users with an environment for communication, which is of great importance irrespective of whether the user is a patient or an ordinary user. This is consistent with the research of Delisle et al [67], which summarized 7 different perceived benefits of participating in rare disease support groups, including giving and receiving emotional support and having a place to speak openly about the disease and one's feelings. Furthermore, membership in online groups can provide those living with long-term conditions with readily available access to self-management and emotional support [68]. The most important positive and negative sentiments were encouragement and worry, indicating that users can get support in OHCs, which will help them overcome difficulties. Negative emotions reflect the worrying situation of patients with albinism and their families. The main issues include a lack of medical-related knowledge, limited amount of national policy on rare diseases, and inferiority caused by the disease. This requires attention from social and medical experts.

Strengthen the Construction

Finally, the construction of OHCs for albinism should be strengthened to better meet the needs of patients. Based on our analysis of the albinism community, the services from OHCs did not meet the users' demand. And this contradiction has gradually intensified. Coincidentally, the situation in other albinism communities in China is also serious. Moon Kids Home [69], a relatively professional platform, is currently the largest OHC for albinism in China. Owing to a lack of management, there is a lot of advertising and spam, preventing the platform from functioning normally. The population of patients is small and geographically scattered [70]. It is therefore difficult to organize effective diagnosis and treatment services. We must be aware of the necessity and urgency of building rare disease OHCs. OHCs facilitate patients' access to health care and increase the availability of medical resources. Relevant medical institutions, companies, and government agencies should establish and maintain professional OHCs in the field of rare diseases, which can be single-species or comprehensive, providing a better community environment for patients. OHCs

can also effectively assist health care providers in collecting patient information. This information assists providers, informaticians, and online health information entrepreneurs in helping patients and caregivers make informed choices [66]. Users of OHCs acquire knowledge and advice related to health risk evaluation, disease prevention and diagnosis, and treatment suggestions from doctors [65]. In addition, patients may provide self-tracking measurements of vital signs and other biological or behavioral parameters that can be transmitted through the internet and allow for richer information for clinical decisions [71].

In developed countries, organizations focused on rare diseases emerged earlier and developed more rapidly. In the field of albinism, there are already some influential organizations, such as the National Organization for Albinism and Hypopigmentation [72], Albinism Fellowship [73], and Albinism Europe, with patients being able to ask for help through the network. Offline care activities are also carried out, but there is still insufficient space to provide free communication. Given China's large population, it is generally believed that the country also has the largest population of people affected by rare diseases [74]. Furthermore, government agencies in China have issued the China's First List, which lists 121 rare diseases to facilitate their management [75]. However, the development gap of relevant domestic forums is obvious. Patients with rare diseases and their families are vulnerable in society and deserve more attention and care.

Implications

The focus of this study is patients with albinism who are easily overlooked and misunderstood by health care providers. OHCs provide the general public with an opportunity to increase their awareness and understanding of the disease. Through topic mining and sentiment analysis, we captured the needs of patients relating to health care, beauty, and making friends. At the same time, we clearly observed obstacles for patients in terms of occupation, education, and social activities, which illustrates the inconvenience caused by physical differences and public discrimination. The role of the albinism community is gradually disintegrating. Obviously, society needs to devote more attention to patients with rare diseases. Relevant health care departments should formulate effective countermeasures based on problems revealed by the results of this study. In addition, this study should also remind us to improve OHCs to satisfy the various needs of patients. We should strengthen psychological counseling via OHCs while improving the living conditions for patients with albinism. Of course, protecting the rights of patients should also be a major priority. All of these require that related agencies, such as medical institutions, companies, and government agencies, establish more professional OHCs for rare diseases based on international experience. In addition, multisector cooperation would allow for the establishment of norms for the creation of OHCs for rare diseases. The research results can only be used as a reference for other rare diseases.

Limitations

Although findings are based on the conducted analysis, there are still several potential limitations that may encourage further research efforts. First, because there are few OHCs for albinism in China, this study has a limited amount of data, which will have a certain effect on the outcome. Due to the limitations of Baidu Tieba, the fields in which to crawl for data have almost no descriptive indicators for the user. Social network analysis only focuses on the mutual connection of users. Second, although the RAs were trained to mark the corpora to ensure the consistency of the labeling results, the topic labeling process was manual, which might introduce bias to the topic evolution. Third, during the labeling process of supervised learning, part of the corpus had both positive and negative emotion expressions. We mainly used its core sentiment for labeling. This process could cause deviations in sentiment polarity to some extent. However, this situation has little impact on the overall distribution, as the corpora collected were mostly short text. Finally, the sentimental polarity for albinism would change over time due to the change in perception or attitude of the Chinese society towards the patients' condition. However, such an evolution was not reflected in our study, which could also lead to bias in the analysis and discussion of the sentimental polarity to some extent.

Conclusions

The combination of topic mining, social network analysis, and sentiment polarity analysis can effectively capture the topics and emotional characteristics of OHC users. This study provides new perspectives for understanding the needs and situations of patients with rare diseases. The albinism community provides a platform for free expression and consultation for Chinese patients with albinism and their families. They have a great demand for medical, inspection, policy, and other related information. Further studies are needed to detect change and the reasons for the sentimental polarity for albinism in OHCs. In addition, research should explore how to strengthen the cooperation of multiple parties to better exert sufficient influence and roles in OHCs. Meanwhile, studies should also be conducted to strengthen the understanding of the social adaptability and psychology of rare disease groups to better learn patient needs.

Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities, HUST (No. 2019WKYXZX011). The authors would like to thank all anonymous reviewers for their valuable comments and input to this research.

Authors' Contributions

QB, the co-first author, designed the study and contributed to the collection of data and writing of the manuscript. LS, the co-first author and corresponding author, designed and conducted the study and finalized the draft manuscript. RE, the third author,

contributed to the writing of the manuscript and final proofreading. ZZ, the fourth author, reviewed the final manuscript. All authors contributed to the preparation and approval of the final accepted version.

Conflicts of Interest

None declared.

Multimedia Appendix 1
The total 36 categories obtained from Latent Dirichlet Allocation model, as well as their merging process.
[DOC File , 85 KB - medinform_v8i5e17813_app1.doc]

References

1. McCafferty BK, Wilk MA, McAllister JT, Stepien KE, Dubis AM, Brilliant MH, et al. Clinical Insights Into Foveal Morphology in Albinism. J Pediatr Ophthalmol Strabismus 2015 May;52(3):167-172 [FREE Full text] [doi: 10.3928/01913913-20150427-06] [Medline: 26053207]
2. Yanes T, Humphreys L, McInerney-Leo A, Biesecker B. Factors Associated with Parental Adaptation to Children with an Undiagnosed Medical Condition. J Genet Couns 2017 Aug;26(4):829-840 [FREE Full text] [doi: 10.1007/s10897-016-0060-9] [Medline: 28039658]
3. Baumbusch J, Mayer S, Sloan-Yip I. Alone in a Crowd? Parents of Children with Rare Diseases' Experiences of Navigating the Healthcare System. J Genet Couns 2019 Feb;28(1):80-90 [FREE Full text] [doi: 10.1007/s10897-018-0294-9] [Medline: 30128673]
4. Pelentsov LJ, Laws TA, Esterman AJ. The supportive care needs of parents caring for a child with a rare disease: A scoping review. Disabil Health J 2015 Oct;8(4):475-491 [FREE Full text] [doi: 10.1016/j.dhjo.2015.03.009] [Medline: 25959710]
5. Dellve L, Samuelsson L, Tallborn A, Fasth A, Hallberg LR. Stress and well-being among parents of children with rare diseases: a prospective intervention study. J Adv Nurs 2006 Feb;53(4):392-402. [doi: 10.1111/j.1365-2648.2006.03736.x] [Medline: 16448482]
6. Dawkins HJ, Draghia-Akli R, Lasko P, Lau LP, Jonker AH, Cutillo CM, International Rare Diseases Research Consortium (IRDiRC). Progress in Rare Diseases Research 2010-2016: An IRDiRC Perspective. Clin Transl Sci 2018 Jan;11(1):11-20 [FREE Full text] [doi: 10.1111/cts.12501] [Medline: 28796411]
7. Abbas A, Vella Szijj J, Azzopardi LM, Serracino Inglott A. Orphan drug policies in different countries. J Pharm Health Serv Res 2019 May 27;10(3):295-302. [doi: 10.1111/jphs.12305]
8. Rodwell C, Aymé S. Rare disease policies to improve care for patients in Europe. Biochim Biophys Acta 2015 Oct;1852(10 Pt B):2329-2335 [FREE Full text] [doi: 10.1016/j.bbadis.2015.02.008] [Medline: 25725454]
9. Dharssi S, Wong-Rieger D, Harold M, Terry S. Review of 11 national policies for rare diseases in the context of key patient needs. Orphanet J Rare Dis 2017 Mar 31;12(1):63 [FREE Full text] [doi: 10.1186/s13023-017-0618-0] [Medline: 28359278]
10. Gomes JDS. [Social identity of people with rare conditions and the lack of diagnosis: contributions based on Hall, Honneth and Jutel]. Cien Saude Colet 2019;24(10):3701-3708 [FREE Full text] [doi: 10.1590/1413-812320182410.12862019] [Medline: 31576999]
11. Marco Leimeister J, Schweizer K, Leimeister S, Krcmar H. Do virtual communities matter for the social support of patients? Info Technology & People 2008 Nov 14;21(4):350-374. [doi: 10.1108/09593840810919671]
12. Huh J, Kwon BC, Kim S, Lee S, Choo J, Kim J, et al. Personas in online health communities. J Biomed Inform 2016 Oct;63:212-225 [FREE Full text] [doi: 10.1016/j.jbi.2016.08.019] [Medline: 27568913]
13. Lu Y, Wu Y, Liu J, Li J, Zhang P. Understanding Health Care Social Media Use From Different Stakeholder Perspectives: A Content Analysis of an Online Health Community. J Med Internet Res 2017 Apr 07;19(4):e109 [FREE Full text] [doi: 10.2196/jmir.7087] [Medline: 28389418]
14. Yan Z, Wang T, Chen Y, Zhang H. Knowledge sharing in online health communities: A social exchange theory perspective. Information & Management 2016 Jul;53(5):643-653. [doi: 10.1016/j.im.2016.02.001]
15. Guo S, Guo X, Fang Y, Vogel D. How Doctors Gain Social and Economic Returns in Online Health-Care Communities: A Professional Capital Perspective. Journal of Management Information Systems 2017 Aug 17;34(2):487-519. [doi: 10.1080/07421222.2017.1334480]
16. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci 2016 Apr;25(2):113-122 [FREE Full text] [doi: 10.1017/S2045796015001067] [Medline: 26744309]
17. Willis E, Royne MB. Online Health Communities and Chronic Disease Self-Management. Health Commun 2017 Mar;32(3):269-278. [doi: 10.1080/10410236.2016.1138278] [Medline: 27218836]
18. Kaur W, Balakrishnan V, Rana O, Sinniah A. Liking, sharing, commenting and reacting on Facebook: User behaviors' impact on sentiment intensity. Telematics and Informatics 2019 Jun;39:25-36. [doi: 10.1016/j.tele.2018.12.005]

19. Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. BMC Med Inform Decis Mak 2018 Jan 05;18(1):2 [FREE Full text] [doi: 10.1186/s12911-017-0579-1] [Medline: 29304788]
20. Brusilovskiy E, Townley G, Snethen G, Salzer MS. Social media use, community participation and psychological well-being among individuals with serious mental illnesses. Computers in Human Behavior 2016 Dec;65:232-240. [doi: 10.1016/j.chb.2016.08.036]
21. Shen L, Wang S, Chen W, Fu Q, Evans R, Lan F, et al. Understanding the Function Constitution and Influence Factors on Communication for the WeChat Official Account of Top Tertiary Hospitals in China: Cross-Sectional Study. J Med Internet Res 2019 Dec 09;21(12):e13025 [FREE Full text] [doi: 10.2196/13025] [Medline: 31815674]
22. Rodrigues RG, das Dores RM, Camilo-Junior CG, Rosa TC. SentiHealth-Cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. Int J Med Inform 2016 Jan;85(1):80-95. [doi: 10.1016/j.ijmedinf.2015.09.007] [Medline: 26514078]
23. Davies W. Insights into rare diseases from social media surveys. Orphanet J Rare Dis 2016 Nov 09;11(1):151 [FREE Full text] [doi: 10.1186/s13023-016-0532-x] [Medline: 27829465]
24. Voigtländer T. Orphan diseases. Why rare diseases need many networks. Monatsschr Kinderheilkd 2012 Sep 5;160(9):863-875. [doi: 10.1007/s00112-012-2668-7]
25. Mártinez-García M, Montoliu L. Albinism in Europe. J Dermatol 2013 May;40(5):319-324. [doi: 10.1111/1346-8138.12170] [Medline: 23668539]
26. Kubasch AS, Meurer M. Oculocutaneous and ocular albinism. Hautarzt 2017 Nov;68(11):867-875. [doi: 10.1007/s00105-017-4061-x] [Medline: 29018889]
27. Grønskov K, Brøndum-Nielsen K, Lorenz B, Preising MN. Clinical utility gene card for: Oculocutaneous albinism. Eur J Hum Genet 2014 Aug;22(8) [FREE Full text] [doi: 10.1038/ejhg.2013.307] [Medline: 24518832]
28. Sun W, Shen Y, Shan S, Han L, Li Y, Zhou Z, et al. Identification of TYR mutations in patients with oculocutaneous albinism. Mol Med Rep 2018 Jun;17(6):8409-8413. [doi: 10.3892/mmr.2018.8881] [Medline: 29658579]
29. George A, Zand D, Hufnagel R, Sharma R, Sergeev Y, Legare J, et al. Biallelic Mutations in MITF Cause Coloboma, Osteopetrosis, Microphthalmia, Macrocephaly, Albinism, and Deafness. Am J Hum Genet 2016 Dec 01;99(6):1388-1394 [FREE Full text] [doi: 10.1016/j.ajhg.2016.11.004] [Medline: 27889061]
30. Kamaraj B, Purohit R. Mutational Analysis on Membrane Associated Transporter Protein (MATP) and Their Structural Consequences in Oculocutaeous Albinism Type 4 (OCA4)-A Molecular Dynamics Approach. J Cell Biochem 2016 Nov;117(11):2608-2619. [doi: 10.1002/jcb.25555] [Medline: 27019209]
31. Fukuda N, Naito S, Masukawa D, Kaneda M, Miyamoto H, Abe T, et al. Expression of ocular albinism 1 (OA1), 3, 4-dihydroxy- L-phenylalanine (DOPA) receptor, in both neuronal and non-neuronal organs. Brain Res 2015 Mar 30;1602:62-74. [doi: 10.1016/j.brainres.2015.01.020] [Medline: 25601010]
32. Wei A, Zang D, Zhang Z, Yang X, Li W. Prenatal genotyping of four common oculocutaneous albinism genes in 51 Chinese families. J Genet Genomics 2015 Jun 20;42(6):279-286. [doi: 10.1016/j.jgg.2015.05.001] [Medline: 26165494]
33. Kruijt CC, de Wit GC, Bergen AA, Florijn RJ, Schalij-Delfos NE, van Genderen MM. The Phenotypic Spectrum of Albinism. Ophthalmology 2018 Dec;125(12):1953-1960. [doi: 10.1016/j.ophtha.2018.08.003] [Medline: 30098354]
34. Kruijt CC, de Wit GC, Talsma HE, Schalij-Delfos NE, van Genderen MM. The Detection Of Misrouting In Albinism: Evaluation of Different VEP Procedures in a Heterogeneous Cohort. Invest Ophthalmol Vis Sci 2019 Sep 03;60(12):3963-3969. [doi: 10.1167/iovs.19-27364] [Medline: 31560370]
35. Thomas MG, Maconachie GD, Sheth V, McLean RJ, Gottlob I. Development and clinical utility of a novel diagnostic nystagmus gene panel using targeted next-generation sequencing. Eur J Hum Genet 2017 Jun;25(6):725-734 [FREE Full text] [doi: 10.1038/ejhg.2017.44] [Medline: 28378818]
36. Brilliant MH. Albinism in Africa: a medical and social emergency. Int Health 2015 Jul;7(4):223-225. [doi: 10.1093/inthealth/ihv039] [Medline: 26063702]
37. Maia M, Volpini BMF, dos Santos GA, Rujula MJP. Quality of life in patients with oculocutaneous albinism. An Bras Dermatol 2015;90(4):513-517 [FREE Full text] [doi: 10.1590/abd1806-4841.20153498] [Medline: 26375220]
38. Wakida-Kusunoki AT. First record of total albinism in southern stingray Dasyatis americana. Rev. biol. mar. oceanogr 2015 Apr;50(1):135-139. [doi: 10.4067/s0718-19572015000100011]
39. Wishkerman A, Boglino A, Darias MJ, Andree KB, Estévez A, Gisbert E. Image analysis-based classification of pigmentation patterns in fish: A case study of pseudo-albinism in Senegalese sole. Aquaculture 2016 Nov;464:303-308. [doi: 10.1016/j.aquaculture.2016.06.040]
40. Albinismbar -Baidu Tieba-here is the harbor of the moon angels and friends. URL: http://tieba.baidu.com/f?kw=%E7%99%BD%E5%8C%96%E7%97%85&fr=index&red_tag=o2761451476 [accessed 2019-04-11]
41. Liu C, Lu X. Analyzing hidden populations online: topic, emotion, and social network of HIV-related users in the largest Chinese online community. BMC Med Inform Decis Mak 2018 Jan 05;18(1):2 [FREE Full text] [doi: 10.1186/s12911-017-0579-1] [Medline: 29304788]
42. Python Software Foundation. Python Release Python 3.7.0 | Python.org. Python Language Reference, version 3.7 URL: https://www.python.org/downloads/release/python-370/ [accessed 2020-04-19]

43. Scrapy | A Fast and Powerful Scraping and Web Crawling Framework. URL: https://scrapy.org/ [accessed 2020-04-19]
44. PyPI. jieba URL: https://pypi.org/project/jieba/ [accessed 2019-05-15]
45. Zhang L, Hall M, Bastola D. Utilizing Twitter data for analysis of chemotherapy. Int J Med Inform 2018 Dec;120:92-100. [doi: 10.1016/j.ijmedinf.2018.10.002] [Medline: 30409350]
46. Printz H, Olsen PA. Theory and practice of acoustic confusability. Computer Speech & Language 2002 Jan;16(1):131-164. [doi: 10.1006/csla.2001.0188]
47. Guo Y, Barnes SJ, Jia Q. Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. Tourism Management 2017 Apr;59:467-483. [doi: 10.1016/j.tourman.2016.09.009]
48. Klakow D, Peters J. Testing the correlation of word error rate and perplexity. Speech Communication 2002 Sep;38(1-2):19-28. [doi: 10.1016/s0167-6393(01)00041-3]
49. Guido S, Mueller AC. Introduction to Machine Learning with Python. Boston, MA: O'Reilly Media; 2016.
50. Shiau W, Dwivedi YK, Yang HS. Co-citation and cluster analyses of extant literature on social networks. International Journal of Information Management 2017 Oct;37(5):390-399. [doi: 10.1016/j.ijinfomgt.2017.04.007]
51. Shen L, Wang S, Dai W, Zhang Z. Detecting the Interdisciplinary Nature and Topic Hotspots of Robotics in Surgery: Social Network Analysis and Bibliometric Study. J Med Internet Res 2019 Mar 26;21(3):e12625 [FREE Full text] [doi: 10.2196/12625] [Medline: 30912752]
52. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. PLoS One 2014;9(6):e98679 [FREE Full text] [doi: 10.1371/journal.pone.0098679] [Medline: 24914678]
53. Kim J, Hastak M. Social network analysis: Characteristics of online social networks after a disaster. International Journal of Information Management 2018 Feb;38(1):86-96. [doi: 10.1016/j.ijinfomgt.2017.08.003]
54. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J. Stat. Mech 2008 Oct 09;2008(10):P10008. [doi: 10.1088/1742-5468/2008/10/P10008]
55. Liu B. Sentiment Analysis: Mining Opinions, Sentiments, Emotions. Cambridge, England: Cambridge University Press; 2015.
56. Faust O, Hagiwara Y, Hong TJ, Lih OS, Acharya UR. Deep learning for healthcare applications based on physiological signals: A review. Comput Methods Programs Biomed 2018 Jul;161:1-13. [doi: 10.1016/j.cmpb.2018.04.005] [Medline: 29852952]
57. Mukhtar N, Khan MA, Chiragh N. Lexicon-based approach outperforms Supervised Machine Learning approach for Urdu Sentiment Analysis in multiple domains. Telematics and Informatics 2018 Dec;35(8):2173-2183. [doi: 10.1016/j.tele.2018.08.003]
58. Fu X, Yang J, Li J, Fang M, Wang H. Lexicon-Enhanced LSTM With Attention for General Sentiment Analysis. IEEE Access 2018;6:71884-71891. [doi: 10.1109/access.2018.2878425]
59. Welcome to HowNet!. URL: http://www.keenage.com/ [accessed 2020-04-19]
60. Hasan A, Moin S, Karim A, Shamshirband S. Machine Learning-Based Sentiment Analysis for Twitter Accounts. MCA 2018 Feb 27;23(1):11. [doi: 10.3390/mca23010011]
61. Tozzi AE, Mingarelli R, Agricola E, Gonfiantini M, Pandolfi E, Carloni E, et al. The internet user profile of Italian families of patients with rare diseases: a web survey. Orphanet J Rare Dis 2013 May 16;8:76 [FREE Full text] [doi: 10.1186/1750-1172-8-76] [Medline: 23680013]
62. Aymé S, Kole A, Groft S. Empowerment of patients: lessons from the rare diseases community. Lancet 2008 Jun 14;371(9629):2048-2051. [doi: 10.1016/S0140-6736(08)60875-2] [Medline: 18555918]
63. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported outcomes as a source of evidence in off-label prescribing: analysis of data from PatientsLikeMe. J Med Internet Res 2011 Jan 21;13(1):e6 [FREE Full text] [doi: 10.2196/jmir.1643] [Medline: 21252034]
64. Gold J, Pedrana AE, Stoove MA, Chang S, Howard S, Asselin J, et al. Developing health promotion interventions on social networking sites: recommendations from The FaceSpace Project. J Med Internet Res 2012 Feb 28;14(1):e30 [FREE Full text] [doi: 10.2196/jmir.1875] [Medline: 22374589]
65. Hajli MN, Sims J, Featherman M, Love PE. Credibility of information in online communities. Journal of Strategic Marketing 2014 May 22;23(3):238-253. [doi: 10.1080/0965254X.2014.920904]
66. Nath C, Huh J, Adupa AK, Jonnalagadda SR. Website Sharing in Online Health Communities: A Descriptive Analysis. J Med Internet Res 2016 Jan 13;18(1):e11 [FREE Full text] [doi: 10.2196/jmir.5237] [Medline: 26764193]
67. Delisle VC, Gumuchian ST, Rice DB, Levis AW, Kloda LA, Körner A, et al. Perceived Benefits and Factors that Influence the Ability to Establish and Maintain Patient Support Groups in Rare Diseases: A Scoping Review. Patient 2017 Jun;10(3):283-293. [doi: 10.1007/s40271-016-0213-9] [Medline: 28004275]
68. Bjarnadottir RI, Millery M, Fleck E, Bakken S. Correlates of online health information-seeking behaviors in a low-income Hispanic community. Inform Health Soc Care 2016 Dec;41(4):341-349 [FREE Full text] [doi: 10.3109/17538157.2015.1064429] [Medline: 26837012]
69. Moon kids home. URL: http://www.albinism.org.cn/ [accessed 2019-05-20]

70. Min R, Zhang X, Fang P, Wang B, Wang H. Health service security of patients with 8 certain rare diseases: evidence from China's national system for health service utilization of patients with healthcare insurance. Orphanet J Rare Dis 2019 Aug 20;14(1):204 [FREE Full text] [doi: 10.1186/s13023-019-1165-7] [Medline: 31429789]
71. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health 2009 Feb;6(2):492-525 [FREE Full text] [doi: 10.3390/ijerph6020492] [Medline: 19440396]
72. National Organization for Albinism and Hypopigmentation. URL: https://www.albinism.org/ [accessed 2019-10-23]
73. Home - Albinism Fellowship UK and Ireland. URL: https://www.albinism.org.uk/ [accessed 2019-10-23]
74. Cui Y, Han J. Defining rare diseases in China. Intractable Rare Dis Res 2017 May;6(2):148-149 [FREE Full text] [doi: 10.5582/irdr.2017.01009] [Medline: 28580219]
75. He J, Tang M, Zhang X, Chen D, Kang Q, Yang Y, et al. Incidence and prevalence of 121 rare diseases in China: Current status and challenges. Intractable Rare Dis Res 2019 May;8(2):89-97 [FREE Full text] [doi: 10.5582/irdr.2019.01066] [Medline: 31218158]

Abbreviations

CNN: convolutional neural network
LDA: latent dirichlet allocation
LSTM: long short-term memory
NB: naive Bayes
OHC: online health community
RAs: research assistants
SVM: support vector machine

Edited by G Eysenbach; submitted 14.01.20; peer-reviewed by T Ndabu, T Muto, V Osadchiy; comments to author 23.02.20; revised version received 05.03.20; accepted 23.03.20; published 29.05.20.

Please cite as:
Bi Q, Shen L, Evans R, Zhang Z, Wang S, Dai W, Liu C
Determining the Topic Evolution and Sentiment Polarity for Albinism in a Chinese Online Health Community: Machine Learning and Social Network Analysis
JMIR Med Inform 2020;8(5):e17813
URL: http://medinform.jmir.org/2020/5/e17813/
doi:10.2196/17813
PMID:32469320

Original Paper

An App Developed for Detecting Nurse Burnouts Using the Convolutional Neural Networks in Microsoft Excel: Population-Based Questionnaire Study

Yi-Lien Lee[1,2], BA; Willy Chou[3,4*], MD; Tsair-Wei Chien[5*], MBA; Po-Hsin Chou[6,7*], MD; Yu-Tsen Yeh[8], MA; Huan-Fang Lee[9], PhD

[1]Department of Medical Affairs, Chi Mei Medical Center, Tainan, Taiwan
[2]Department of Information Management and Institute of Healthcare Information Management, National Chung Cheng University, Chayi, Taiwan
[3]Department of Physical Medicine and Rehabilitation, Chiali Chi Mei Hospital, Chi Mei Medical Groups, Tainan, Taiwan
[4]Department of Physical Medicine and Rehabilitation, Chung Shan Medical University, Taichun, Taiwan
[5]Department of Medical Research, Chi Mei Medical Center, Chi Mei Medical Groups, Tainan, Taiwan
[6]Department of Orthopedics and Traumatology, Taipei Veterans General Hospital, Taipei, Taiwan
[7]School of Medicine, National Yang-Ming University, Taipei, Taiwan
[8]Medical School, St George's, University of London, London, United Kingdom
[9]Department of Nursing, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[*]these authors contributed equally

Corresponding Author:
Huan-Fang Lee, PhD
Department of Nursing
College of Medicine
National Cheng Kung University
988 Chung Hwa Road
Yung Kung District
Tainan
Taiwan
Phone: 886 62812811
Email: Eamonn0330@gmail.com

Abstract

Background: Burnout (BO), a critical syndrome particularly for nurses in health care settings, substantially affects their physical and psychological status, the institute's well-being, and indirectly, patient outcomes. However, objectively classifying BO levels has not been defined and noticed in the literature.

Objective: The aim of this study is to build a model using the convolutional neural network (CNN) to develop an app for automatic detection and classification of nurse BO using the Maslach Burnout Inventory–Human Services Survey (MBI-HSS) to help assess nurse BO at an earlier stage.

Methods: We recruited 1002 nurses working in a medical center in Taiwan to complete the Chinese version of the 20-item MBI-HSS in August 2016. The k-mean and CNN were used as unsupervised and supervised learnings for dividing nurses into two classes (n=531 and n=471 of suspicious BO+ and BO−, respectively) and building a BO predictive model to estimate 38 parameters. Data were separated into training and testing sets in a proportion 70%:30%, and the former was used to predict the latter. We calculated the sensitivity, specificity, and receiver operating characteristic curve (area under the curve) across studies for comparison. An app predicting respondent BO was developed involving the model's 38 estimated parameters for a website assessment.

Results: We observed that (1) the 20-item model yields a higher accuracy rate (0.95) with an area under the curve of 0.97 (95% CI 0.94-0.95) based on the 1002 cases, (2) the scheme named matching personal response to adapt for the correct classification in model drives the prior model's predictive accuracy at 100%, (3) the 700-case training set with 0.96 accuracy predicts the 302-case testing set reaching an accuracy of 0.91, and (4) an available MBI-HSS app for nurses predicting BO was successfully developed and demonstrated in this study.

Conclusions: The 20-item model with the 38 parameters estimated by using CNN for improving the accuracy of nurse BO has been particularly demonstrated in Excel (Microsoft Corp). An app developed for helping nurses to self-assess job BO at an early stage is required for application in the future.

(JMIR Med Inform 2020;8(5):e16528) doi:10.2196/16528

KEYWORDS

nurse burnout; MBI-HSS Chinese version; receiver operating characteristic curve; convolutional neural network; Lz person fit statistic

Introduction

Burnout in the Workplace

Burnout (BO) is a critical syndrome and problem in high-tech service-oriented societies, particularly for nurses in health care settings [1-4]. Many studies [5-11] reported that BO influences an employee's physical and psychological status [5-7], the organizational well-being [8-11], and patient quality-of-care outcomes [6,10].

One of the most popular BO inventories is the Maslach Burnout Inventory–Human Services Survey (MBI-HSS) [12,13]. More than 1898 articles were found by searching the keywords "Maslach" and "burnout" on September 23, 2019. BO is defined by Maslach [12,13] as a syndrome of emotional exhaustion, reduced personal accomplishment (PA), and depersonalization that frequently occurs in individuals who work in people-related jobs, such health care and educational.

Maslach Burnout Inventory–Human Services Survey

The MBI-HSS [13] has been widely applied to measure individual BO in numerous workplaces [4,11,14-16]. The original MBI-HSS is a 22-item inventory with a 7-point scale (from never=0 to every day=6) to measure BO for workers in a recent week [13]. The three BO subscales comprise 9 items for emotional exhaustion, 8 items for personal accomplishment, and 5 items for depersonalization. Despite the survey being popularly used in social science, the cutting point for determining BO substantially differs between cultures and health care settings [15,17-21]. Accordingly, Maslach et al [22] suggested that BO levels (low, moderate, and high) had different cutting points in different countries and areas. Schaufeli and Van Dierendonck [23] suggested having common cutting points to compare BO levels among countries and areas.

Maslach and Jackson [13] suggested that the cutting points be set at 54 for emotional exhaustion, 48 for personal accomplishment, and 30 for depersonalization using subscale scores for measurement. Schaufeli and Van Dierendonck [23] were critical of the fact that the scheme for determining BO levels was arbitrary based on the three groups that contained an equal number of sample sizes [24]. Although Maslach and Jackson [13] also suggested having valid criteria that can be used for classifying BO levels, no such reasonable and viable scheme has been accepted by practitioners in the past.

Convolutional Neural Network

Convolutional neural network (CNN) has had the greatest impact within the field of health informatics [25]. Its architecture can be described as an interleaved set of feedforward layers implementing convolutional filters followed by reduction, rectification, or pooling layers [26-28]. For each layer, the CNN creates a high-level abstract feature. Whether the CNN, a famous deep learning method, can improve the prediction accuracy (up to 7.14%) [28] on nurse BO classification is worthy of study.

Online Classification Using Smartphones is Required

As with all forms of web-based technology, advances in mobile health communication technology are rapidly increasing [29]. Until now, there has been no app for smartphones to classify nurse BO levels. If the CNN BO model's parameters have been estimated by the CNN algorithm, the classification of nurse BO by responding to the MBI-HSS can alert individual nurses more accurately and warn them to alleviate their mental strain before it becomes a serious BO problem.

Study Aims

The aims of our study are to (1) estimate the model's parameters using CNN based on nurse responses to the MBI-HSS and (2) design an app for smartphones based on a website assessment of nurse BO.

Methods

Data Source

Study Sample and Demographic Data

If the confidence level and intervals were set at 0.05 and ±5% and applied to the population of 1850 registered nurses in a hospital, 318 participants are required for the sample size [30]. We estimated the rate of refusal to respond will reach 40%. The minimum number for the study sample size will be 540 (318/[1–0.4]).

We delivered 40 copies each of the MBI-HSS BO survey to 32 nursing units. A sample of 1255 registered nurses with at least 1 month experience in the Chi Mei Medical Center (Taiwan) was randomly selected to complete the Chinese version of the 20-item MBI-HSS [3] in August 2016. A total of 1002 participants were eligible, for a return rate of 79.9%.

Featured Variables

Featured variables consist of the 20 items (called the 20-item model in which the response in the subscale of reduced personal accomplishment has been reversed to be the higher score denoting the more serious BO problem) on the classification of nurse BO levels (ie, suspicious BO+ and BO−). The 1002 participants were split into training and testing sets in a proportion (70%:30%), and the former was used to predict the latter. The data are shown in Multimedia Appendix 1. This study

was approved and monitored by the Chi Mei Medical Center institutional review board (10704-003). All hospital and study participant identifiers were stripped.

Unsupervised and Supervised Learnings

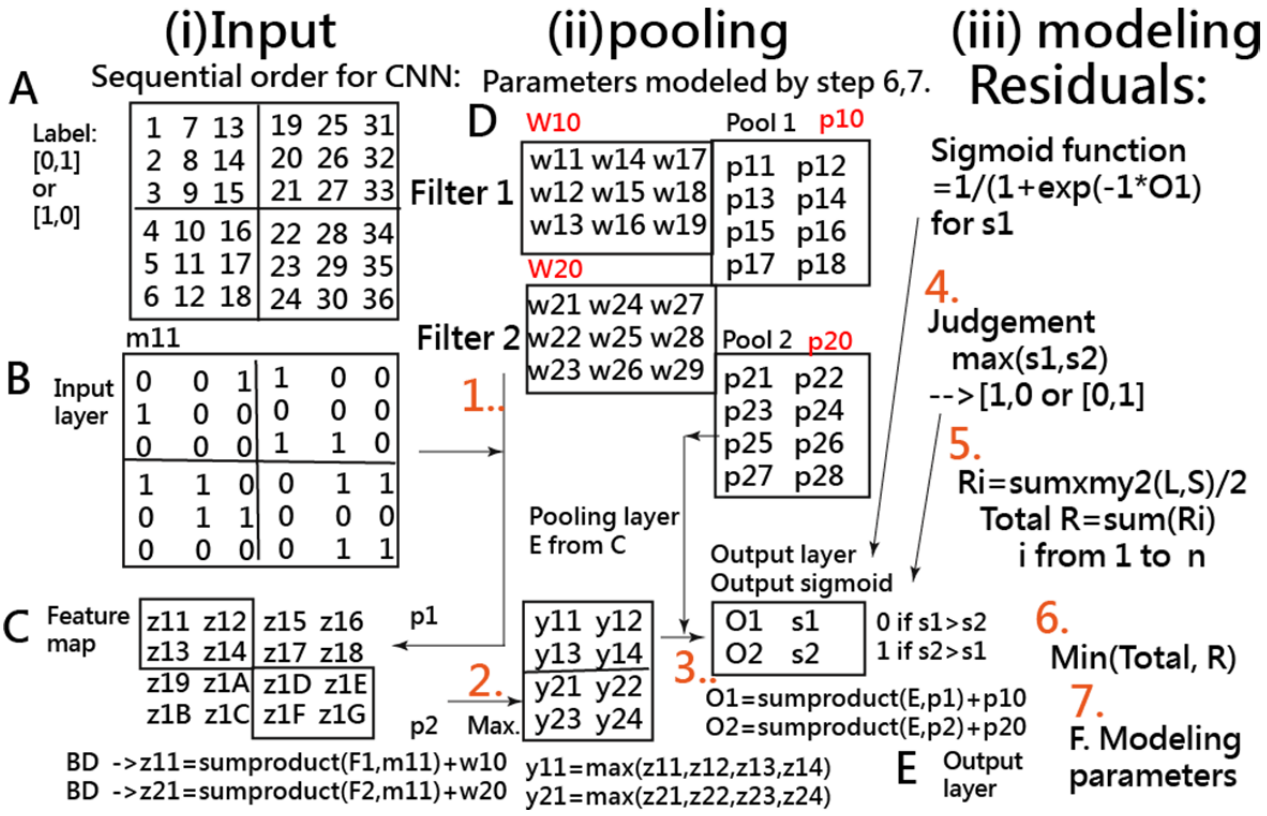
Unsupervised learning indicates agnostic aggregation of unlabeled data sets yielding groups or clusters of entities with shared similarities that may be unknown prior to the analysis step [31,32] (eg, clustering dimensionality reduction using principle component analysis or k-mean clustering). The k-mean clustering aims to partition n observations into k clusters, in which each observation belongs to the cluster with the nearest mean [33]. In contrast, supervised learning employs labeled training data sets (labeled/supervised by subject experts or by the objective k-mean clustering) to yield a qualitative or quantitative output [31,34].

In this study, the k-mean was used as unsupervised learning for clustering participants into two classes (n=531 and n=471 for suspicious BO+ and BO−, respectively). CNN was applied as supervised learning to build a BO prediction model for estimating the 38 parameters.

Convolutional Neural Network Applied in This Study

CNN is a variant of the standard multilayer perceptron, especially used for pattern recognition compared with conventional approaches [35] due to its capability in reducing the dimension of data, extracting the feature sequentially, and classifying one structure of the network [36]. The basic CNN model was inspired in 1962 from the visual cortex proposed by Hubel and Wiesel [35]. For simplifying the CNN concept and process, we present it in Figure 1. Detailed information on interpretation is provided in Multimedia Appendix 2.

Figure 1. Interpretation of the convolutional neural network algorithm.



Tasks for Performing Convolutional Neural Network

Task 1: Comparison of Prediction Accuracies in Two Modes

Two sets of featured variables (ie, 20 with the traditional accurate rate and 100% rate) on 1002 cases were mirrored to compare the prediction accuracies (eg, sensitivity, specificity, and receiver operating characteristic (ROC) curve [area under the curve, AUC]) using the CNN algorithm.

In contrast to the traditionally predictive method, we use the known responses and their corresponding labels (ie, suspicious BO+ or BO−) to build a model for predicting the unknown label of the specific responses. The reason for reaching a 100% accuracy rate on the known responses and their corresponding labels in the training set is to avoid letting the CNN fail in the classification of the known responses in the future. A scheme named matching personal response scheme to adapt for the correct classification in the model (MPRSA) is designed for driving the model's accuracy toward 100%. The way we applied the MPRSA is presented for achieving this 100% goal if the same response string is encountered in the future: the MPRSA is regarding the original responses (eg, the 20-item string coded as 9223372036854775807) that are linked to the correct label in the validation or testing set through which all cases in the training set would reach a 100% accuracy rate if the cases are present in the testing set.

Task 2: Validation Compared With the Training and Testing Sets

The 1002 cases were split into training and testing sets in a proportion of 70%:30%, and the former was used to predict the latter. The accuracy rates in these two sets were compared.

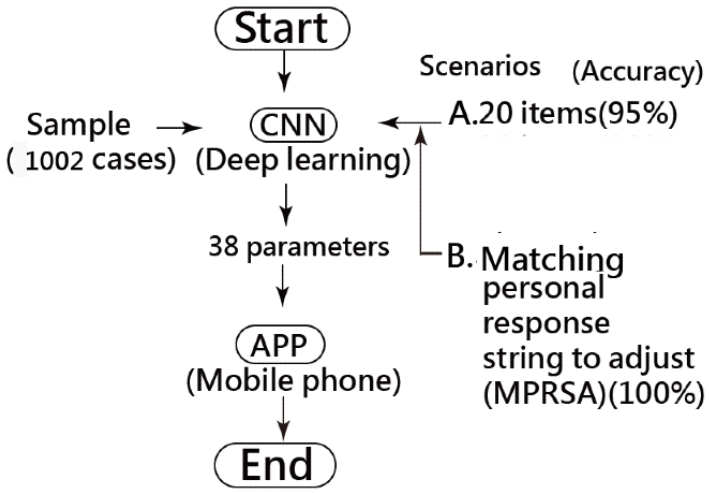
Task 3: App Detecting Burnout for a Web-Based Assessment

A 20-item self-assessment app using participant mobile phones was designed to predict nurse BO using the CNN algorithm and the model parameters [37]. The resulting classification appears on smartphones. The visual representation with binary (BO– and BO+) category probabilities is shown on a dashboard using Google Maps to display.

Statistical Tools and Data Analysis

MedCalc 9.5.0.0 for Windows (MedCalc Software) was used to calculate the sensitivity, specificity, and corresponding AUC using logistic regression when the observed labels (ie, 0 for BO– and 1 for BO+) and the predicted probabilities (ie, the continuous variable in step 3 calculated by the sigmoid function in the output layer in Figure 1) were applied. A visual representation displaying the classification effect is plotted using two curves (ie, one from the left-bottom to the right-top corner denotes the success [BO+] feature and another from the left-top corner to the right-bottom side as the failure attribute). The study flowchart and the CNN modeling process are shown in Figure 2 and Multimedia Appendix 2, respectively.

Figure 2. Study flowchart. CNN: convolutional neural network; MPRSA: matching personal response scheme to adapt for the correct classification in the model.



Results

Demographic Data of Participants

The demographic data of the nurses are shown in Table 1. We can see that females accounted for 93.1% (933/1002) of the participants. Most participants had a bachelor's (university) degree (892/1002, 89.0%). The single accounted for 59.5% (596/1002), and the married (399/1002, 39.8%). Among the nurses, 37.3% (37/1002) had work experience outside the study hospital, while 62.5% (627/1002) had none.

The highest in nurse hierarchy is N (132/1002, 13.2%), followed by N1 (134/1002, 13.4%), N2 (272/1002, 27.1%), N3 (248/1002, 24.8%), and N4 (215/1002, 21.5%). The top two job titles are nurse (797/1002, 79.5%) and leader (149/1002, 14.9%).

The average age for the sample is 32.6 (SD 7.2) years, ranging from 23 to 56. The average work experience in other hospitals reaches 15.1 (SD 28.5) months.

The workload in terms of the number of patients cared for in a week by each nurse averages 11 (SD 19.1). The mean for non-care affairs in a week reaches 4 hours (SD 5.8). The mean of nursing care is 9 (SD 2.7) hours per week. The average number of a patient cared for is 9 (SD 12.1).

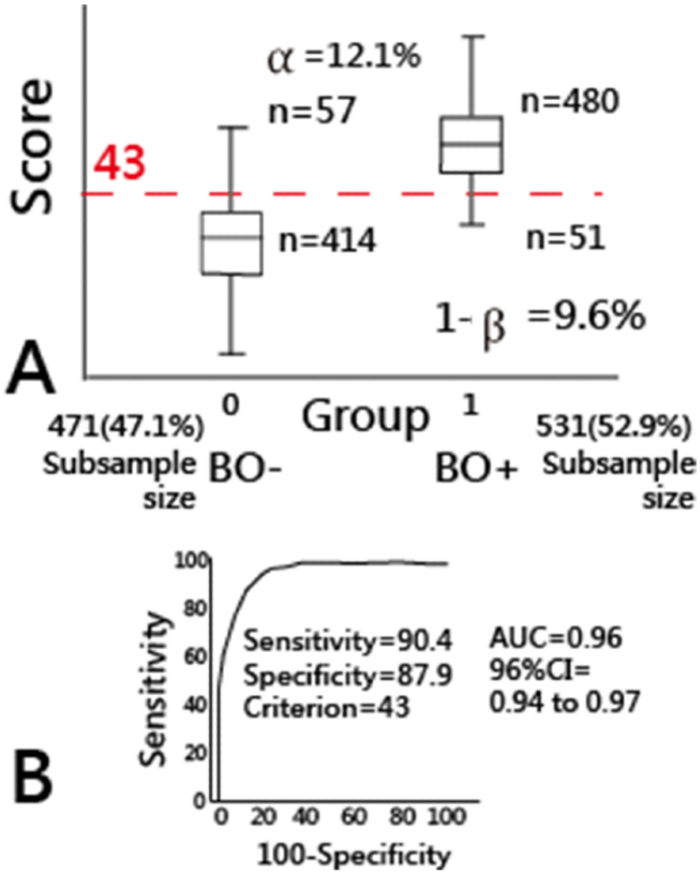
Table 1. Demographic data of the study sample.

Variable and type	Value
Gender, n (%)	
Male	69 (6.9)
Female	933 (93.1)
Education, n (%)	
Less than university	46 (4.6)
University	892 (89.0)
Graduate school	64 (6.4)
Marital status, n (%)	
Single	596 (59.5)
Married	399 (39.8)
Divorced	7 (0.7)
Work tenure, n (%)	
Without	627 (62.6)
With	375 (37.4)
Nurse hierarchy, n (%)	
N (<1 year experience)	133 (13.3)
N1 (Fundamentals of Nursing)	134 (13.4)
N2 (Critical Care in Nursing)	272 (27.1)
N3 (Holistic Care and Teaching)	248 (24.8)
N4 (Specialist Nursing and Research)	215 (21.5)
Job title, n (%)	
Nurse	798 (79.6)
Leader	147 (14.7)
Assistant head nurse	30 (3.0)
Head nurse	27 (2.7)
Age, mean (SD), range	32.6 (7.2), 23-56
Work experience outside hospital (month), mean (SD), range	15.1 (28.5), 0-180
Average hours spent in non-care affairs per week, mean (SD), range	3.9 (5.8), 0-60
Average weekly hours spent in nursing care, mean (SD), range	9.2 (2.9), 1.5-70
Average daily patient care, mean (SD), range	9.5 (12.1), 0-120

Unsupervised Learnings Using the K-Mean Clustering

A visual representation displaying the classification effect is plotted using the box plot (Figure 3). We can see a smaller number of cases with suspicious BO– having a higher total score, and a smaller number of cases are misclassified as BO+ (12.1%) and BO− (9.6%). In contrast, the sensitivity and specificity are 90.4% and 87.9%, respectively. The cutting point is set at 43 with an AUC 0.96 (bottom, Figure 3) if the unsupervised learning approach is applied.

Figure 3. Two study groups divided by the k-mean algorithm (A) and receiver operating characteristic curve (B).



Tasks to Compare the Accuracy Between Modes

Comparison of Prediction Accuracies in Two Modes

The 20-item model yields a higher accuracy rate (0.95) with an AUC 0.98 (95% CI 0.97-1.00) higher than that of the 20-item model with an accuracy of 0.95 and an AUC 0.97 (95% CI 0.96-0.99) based on the 1002 cases.

The MPRSA applied to the bottom pattern in Table 2 drives the model's accuracy at 100%.

Validation Compared With the Training and Testing Sets

The 700-case training set with 0.96 accuracies predicts the 302-case testing set reaching an accuracy of 0.91 (Table 3).

Table 2. Three scenarios applied to convolutional neural network for the prediction of nurse burnout (n=1002).

Sample	True condition			
	BO+[a]	BO–[b]	BO+/row #	BO–/row #
Scenario A (only 20 items)				
Positive	507	26	0.95	0.05
Negative	24	445	0.05	0.95
Scenario B (Scenario A and MPRSA[c]) training				
Positive	531	0	1.00	0
Negative	0	471	0	1.00

[a]BO+: suspicious for burnout.

[b]BO–: not suspicious for burnout.

[c]MPRSA: matching personal response scheme to adapt for the correct classification.

Table 3. Training and testing effects.

Sample	True condition			
	BO+[a]	BO–[b]	BO+/row #	BO–/row #
Scenario A (20 items) training, n=700				
Positive	362	15	0.96	0.04
Negative	10	313	0.03	0.97
Scenario B (20 items) testing, n=302				
Positive	147	16	0.90	0.10
Negative	11	128	0.08	0.92

[a]BO+: suspicious for burnout.

[b]BO–: not suspicious for burnout.

App Detecting Burnout for a Web-Based Assessment

An MBI-HSS app for nurses predicting BO was developed (Figure 4). Interested readers are invited to scan the QR code to practice the MBI-HSS app on their own. It is worth noting that all 38 model parameters are embedded in the 20-item CNN model for classification of either suspicious BO+ or BO− once all 20 items have been responded to.

One resulting example is present in Figure 5, from which we can see that the BO– with a high probability (0.99) is shown on the curve of the failure from the left-top to the right-bottom corner. The sum of both probabilities (ie, BO+ and BO–) equals 1.0. The odds can be computed by the formula (p/[1–p]=0.01/0.99=0.01), indicating the nurse with an extremely low probability or tendency toward BO+.

Figure 4. Screenshot of the mobile phone app.

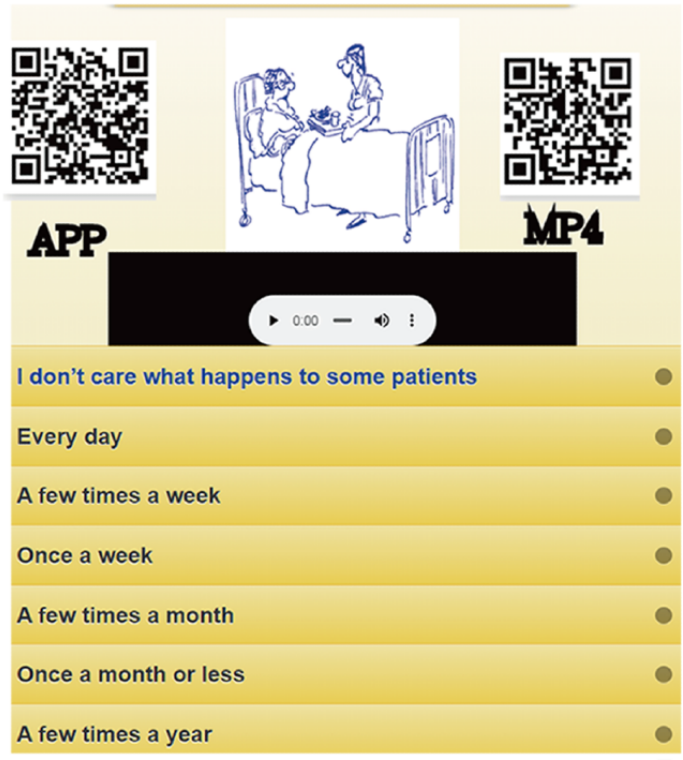
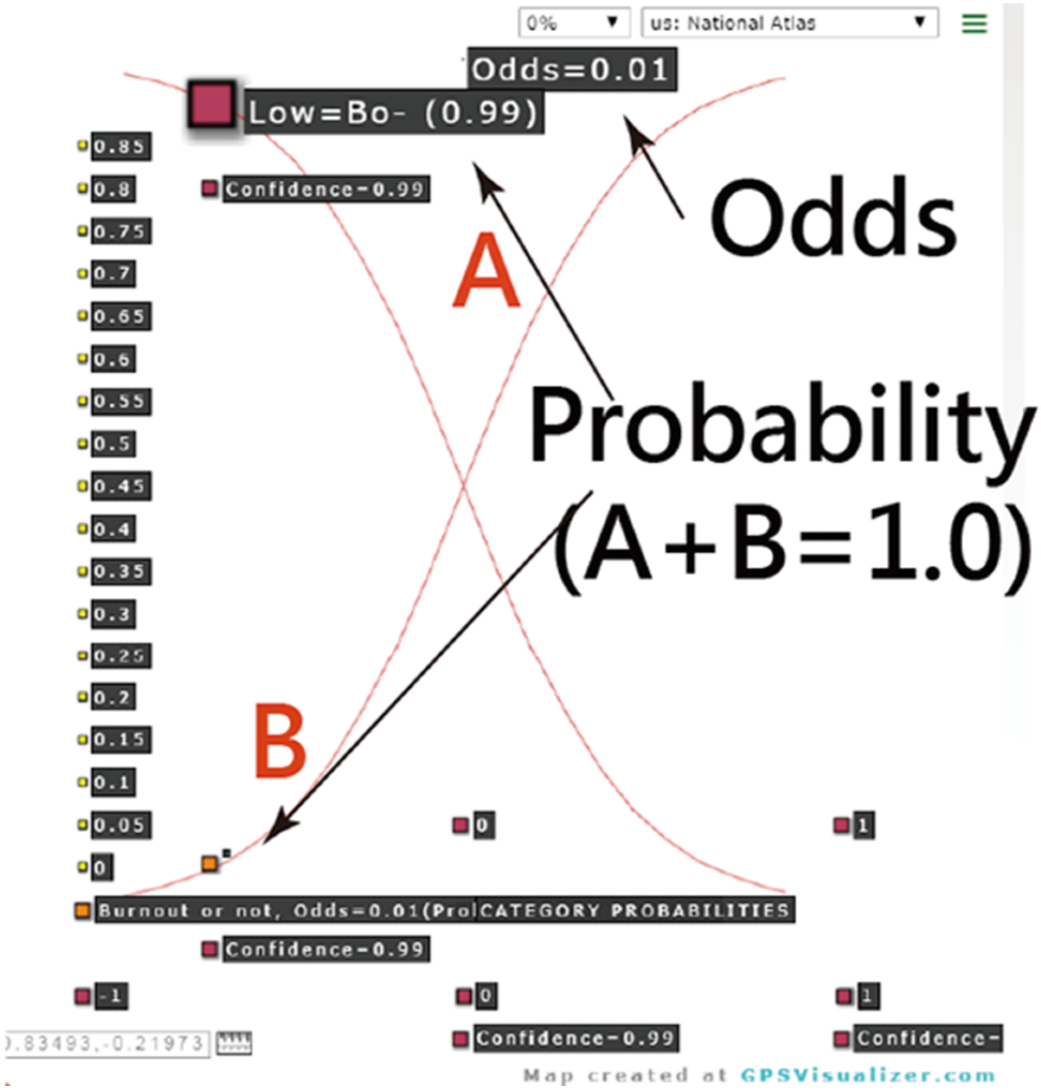


Figure 5. The result of assessing nurse burnout.



Discussion

Principal Findings

We observed that (1) the 20-item model yields a higher accuracy rate (0.95; AUC 0.97, 95% CI 0.94-0.95), (2) the MPRSA drives the model's prior accuracy at 100%, (3) the 700-case training set with 0.96 accuracy predicts the 302-case testing set reaching an accuracy of 0.91, and (4) the MBI-HSS app for nurses predicting BO has been developed and demonstrated.

The MBI-HSS is the most widely used tool for measuring BO in the world [11,14-16]. More than 1898 articles were found by searching the keywords "Maslach" and "burnout" on September 23, 2019. However, none provided an acceptable scheme to classify BO levels (ie, BO+ and BO− or low, moderate, and high).

Maslach and Jackson [13] provided a cutting point scheme (ie, 54 for emotional exhaustion, 48 for personal accomplishment, and 30 for depersonalization; around 40=132/3*20/22) approximately equal to 43 (based on 20 items using subscale total scores; see Figure 3) in this study. Although Schaufeli and Van Dierendonck [23] doubted that the cutting points proposed by Maslach and Jackson [13] were arbitrary only on the assumption of an equal sample size across the levels (ie, high, moderate, and low), our cutting point at 43 is derived through the k-mean clustering.

However, no matter which cutting point scheme is applied, that of Maslach and Jackson [13] or this study (eg, in Figure 3), misclassifications must exist due to their type I (α) and II ($1-\beta$) errors. In contrast, the CNN predictive model combined with the MPRSA mentioned in Methods regarding task 1 (100% accuracy rate is required) can minimize the type I and II errors toward zero (eg, Table 2), which is one of the features of this study.

Implications and Future Work

CNN can improve prediction accuracy (up to 7.14%) [28]. In this study, sensitivity and specificity have been improved. So far, we have not seen anyone using the CNN approach to predict nurse BO in the literature, which is a breakthrough, and the first feature of this study.

Over 708 articles have been found using the keyword "convolutional neural network" (Title) searched in PubMed Central on September 23, 2019. None used Microsoft Excel to perform the CNN. The interpretations for the CNN concept and process or the parameter estimations are shown in Figure 1,

Multimedia Appendix 2 and 3, and in the app [38], which is the second feature of this study.

Using Microsoft Excel to perform CNN is the third feature of this study (Multimedia Appendix 1), which was rarely seen applicable in the literature.

Because the principle for concerning more with the vital few and less with the trivial numerous is emphasized in the quality control process, we propose the MPRSA as the fourth feature. We incorporated the original responses into the model to let the label be correctly classified by the CNN, through which all cases with a false prediction in the training set would be adjusted as a true prediction, reaching a 100% accuracy rate if the cases reoccur in the testing set.

Furthermore, the curves of category probabilities based on the Rasch rating scale model [39] are shown in Figure 4. The binary categories (eg, success and failure on an assessment in the psychometric field) have been applied in health-related outcomes [40-44]. However, none provided the animation-type dashboard showing on Google Maps, as we did in Figure 4.

Strengths

It is easy to set up the nurse BO online assessment if the designer uploads relevant and appropriate audio and visual files to the corresponding questions of the database. We applied the CNN algorithm along with the model's parameters to design the routine on an app that is used to detect BO risk for nurses in hospitals (Figure 4), which has never been seen before for the MBI-HSS [13] implemented on mobile phones.

As with all forms of web-based technology, advances in health communication technology are rapidly emerging [29]. Mobile online BO assessment is promising and worth considering in many fields of health assessment. An online BO assessment (Figure 4) can be applied to inform examinees quickly about when and whether they should take actions or follow up to see a psychiatrist and how to improve their behaviors and attitudes given that their lifestyle is not changed [4]. The online BO assessment is promising, and it is worth using for promoting nurses' health literacy by using the animation-type assessment on smartphones. Interested readers are recommended to scan the QR codes on Figure 4, one for the app and another for the MP4, and see the details about responding to questions and the real experience on answering the 20-item MBI-HSS for a website assessment.

The CNN module on Microsoft Excel is unique and innovative (Multimedia Appendix 1). Users who are not familiar with the CNN software (eg, Python) can apply our Excel Visual Basic for Applications module to conduct CNN-related research in the future. The module is not limited to the binary classification. The multiclassification module can be done by adding the layers on CNN. That is, two categories require two input layers and two pooling layers. Similarly, three categories need three input layers and three pooling layers (Figure 1 and Multimedia Appendix 1 and 2). Any other types of self-assessment, such as work bullying, depression, and dengue fever, can apply the CNN model to predict and classify the levels of harmfulness and disease in the future.

Limitations and Suggestions

Our study has some limitations. First, although the psychometric properties of the 20-item MBI-HSS have been validated for measuring nurse BO in Taiwan [3] after removing item 14 (I feel I am working too hard on my job) and item 22 (I feel patients blame me for some of their problems), there is no evidence that supports that the 20-item MBI-HSS is suitable for nurses in other regions. We recommend additional studies using their own k-mean algorithm and the CNN model to estimate the parameters and see the difference (eg, the cutting point at 43 in Figure 3).

Second, we have not discussed any improvement in predictive accuracy. For instance, whether other featured variables (eg, the mean, SD, and Lz index [44,45]) applied to the CNN model can increase the accurate rate is worthy of further study. Future studies are needed to look for other variables that can improve the power of the model prediction.

Third, the study was based on previously published [3] research using the 20-item MBI-HSS. All of the data were sampled from similar health care settings. If any environment or condition is changed (eg, other professionals or workplaces), the result (eg, the model's parameters) must be different from this study.

Fourth, the MBI-HSS is a three-dimensional construct. Usually, the item difficulties should be first calibrated by using the Rasch ConQuest software [46]. The CNN model [47] can ignore the issue of dimensionality and gain a favorable prediction effect that should be verified and ensured in the future.

Finally, the study sample was taken from Taiwanese data in a nurse survey. The model parameters estimated for the MBI-HSS Chinese version are only suitable for the Chinese (particularly for Taiwanese) society in health care settings. Generalizing these BO assessment findings (eg, the cutting point at around 43; see Figure 3) might be somewhat limited and constrained because the sample merely consisted of nurses working for inpatients. Additional studies are needed to reexamine whether the psychometric properties of the BO assessment are similar to that of other worksites in (or out of) a hospital.

Conclusion

We illustrate features and contributions in this study: (1) CNN performed in Microsoft Excel, (2) MPRSA applied to increase the model's prior prediction accuracy, (3) an online app demonstrated to display results using a visual dashboard on Google Maps, and (4) the category probability curves based on Rasch rating scale model first observed in the CNN prediction model. The novelty of the app with the CNN algorithm improves the predictive accuracy of nurse BO. It is expected to help nurses self-assess job BO at an early stage in the future.

Authors' Contributions

YLL conceived and designed the study, WC and PHC performed the statistical analyses, and YTY was in charge of recruiting study participants. TWC helped design the study, collected information, and interpreted data. HFL monitored the research. All authors read and approved the final article.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Study dataset.
[XLSX File (Microsoft Excel File), 132 KB - medinform_v8i5e16528_app1.xlsx]

Multimedia Appendix 2
Convolutional neural network to interpret Figure 1.
[DOCX File , 81 KB - medinform_v8i5e16528_app2.docx]

Multimedia Appendix 3
Mp4 for convolutional neural network performed in Excel.
[TXT File , 0 KB - medinform_v8i5e16528_app3.txt]

References

1. Abushaikha L, Saca Hazboun H. Job satisfaction and burnout among Palestinian nurses. East Mediterr Health J 2009 Jan 01;15(1):190-197. [doi: 10.26719/2009.15.1.190]
2. Demir A, Ulusoy M, Ulusoy MF. Investigation of factors influencing burnout levels in the professional and private lives of nurses. International Journal of Nursing Studies 2003 Nov;40(8):807-827. [doi: 10.1016/s0020-7489(03)00077-4]
3. Lee H, Chien T, Yen M. Examining factor structure of Maslach burnout inventory among nurses in Taiwan. J Nurs Manag 2013;21:648-656. [doi: 10.1111/j.1365-2834.2012.01427.x] [Medline: 23410056]
4. Lee H. Determining cutting points of the Maslach Burnout Inventory for nurses to measure their level of burnout online. History Res 2017;5(1):1-8. [doi: 10.11648/j.history.20170501.11]
5. Hsu H, Chen S, Yu H, Lou J. Job stress, achievement motivation and occupational burnout among male nurses. J Adv Nurs 2010 Jul;66(7):1592-1601. [doi: 10.1111/j.1365-2648.2010.05323.x] [Medline: 20492017]
6. Spence Laschinger HK, Leiter MP. The impact of nursing work environments on patient safety outcomes: the mediating role of burnout/engagement. J Nurs Adm 2006 May;36(5):259-267. [doi: 10.1097/00005110-200605000-00019] [Medline: 16705307]
7. Trinkoff A, Geiger-Brown J, Brady B, Lipscomb J, Muntaner C. How long and how much are nurses now working? Am J Nurs 2006 Apr;106(4):60-71. [doi: 10.1097/00000446-200604000-00030] [Medline: 16575241]
8. Alacacioglu A, Yavuzsen T, Dirioz M, Oztop I, Yilmaz U. Burnout in nurses and physicians working at an oncology department. Psychooncology 2009 May;18(5):543-548. [doi: 10.1002/pon.1432] [Medline: 18942658]
9. Garrett C. The effect of nurse staffing patterns on medical errors and nurse burnout. AORN J 2008 Jun;87(6):1191-1204. [doi: 10.1016/j.aorn.2008.01.022] [Medline: 18549833]
10. Halbesleben JRB, Wakefield BJ, Wakefield DS, Cooper LB. Nurse burnout and patient safety outcomes: nurse safety perception versus reporting behavior. West J Nurs Res 2008 Aug;30(5):560-577. [doi: 10.1177/0193945907311322] [Medline: 18187408]
11. Spence Laschinger HK, Leiter M, Day A, Gilin D. Workplace empowerment, incivility, and burnout: impact on staff nurse recruitment and retention outcomes. J Nurs Manag 2009 Apr;17(3):302-311. [doi: 10.1111/j.1365-2834.2009.00999.x] [Medline: 19426367]
12. Maslach C. Burnout: The Cost of Caring. Englewood Cliffs: Prentice-Hall; 1982.
13. Maslach C, Jackson S. Maslach Burnout Inventory Manual, 2nd Edition. Palo Alto: Consulting Psychologists Press; 1986.
14. Li XM, Liu YJ. [Job stressors and burnout among staff nurses]. Chin J Nurs 2000;35(11):645.
15. Lin F, St John W, McVeigh C. Burnout among hospital nurses in China. J Nurs Manag 2009 Apr;17(3):294-301. [doi: 10.1111/j.1365-2834.2008.00914.x] [Medline: 21456318]
16. Tourangeau A, Cummings G, Cranley L, Ferron E, Harvey S. Determinants of hospital nurse intention to remain employed: broadening our understanding. J Adv Nurs 2010 Jan;66(1):22-32 [FREE Full text] [doi: 10.1111/j.1365-2648.2009.05190.x] [Medline: 20423434]
17. Beckstead JW. Confirmatory factor analysis of the Maslach Burnout Inventory among Florida nurses. Int J Nurs Stud 2002 Nov;39(8):785-792. [doi: 10.1016/s0020-7489(02)00012-3] [Medline: 12379296]

18. Kanste O, Miettunen J, Kyngäs H. Factor structure of the Maslach Burnout Inventory among Finnish nursing staff. Nurs Health Sci 2006 Dec;8(4):201-207. [doi: 10.1111/j.1442-2018.2006.00283.x] [Medline: 17081145]
19. Poghosyan L, Aiken LH, Sloane DM. Factor structure of the Maslach burnout inventory: an analysis of data from large scale cross-sectional surveys of nurses from eight countries. Int J Nurs Stud 2009 Jul;46(7):894-902 [FREE Full text] [doi: 10.1016/j.ijnurstu.2009.03.004] [Medline: 19362309]
20. Vanheule S, Rosseel Y, Vlerick P. The factorial validity and measurement invariance of the Maslach Burnout Inventory for human services. Stress Health 2007 Apr;23(2):87-91. [doi: 10.1002/smi.1124]
21. Worley JA, Vassar M, Wheeler DL, Barnes LLB. Factor structure of scores from the Maslach Burnout Inventory. Educat Psychol Measure 2008 Feb 05;68(5):797-823. [doi: 10.1177/0013164408315268]
22. Maslach C, Schaufeli WB, Leiter MP. Job burnout. Annu Rev Psychol 2001;52:397-422. [doi: 10.1146/annurev.psych.52.1.397] [Medline: 11148311]
23. Schaufeli WB, Van Dierendonck D. A cautionary note about the cross-national and clinical validity of cut-off points for the Maslach Burnout Inventory. Psychol Rep 1995 Jun;76(3 Pt 2):1083-1090. [doi: 10.2466/pr0.1995.76.3c.1083] [Medline: 7480470]
24. Golembiewski R, Deckard G, Roundtree B. The stability of burnout assignment: measurement properties of the phase model. J Health Hum Serv Admin 1989;12:63-78.
25. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/JBHI.2016.2636665] [Medline: 28055930]
26. Tobore I, Li J, Yuhang L, Al-Handarish Y, Kandwal A, Nie Z, et al. Deep learning intervention for health care challenges: some biomedical domain considerations. JMIR Mhealth Uhealth 2019 Aug 02;7(8):e11966 [FREE Full text] [doi: 10.2196/11966] [Medline: 31376272]
27. Kwon S, Hong J, Choi E, Lee E, Hostallero DE, Kang WJ, et al. Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study. JMIR Mhealth Uhealth 2019 Jun 06;7(6):e12770 [FREE Full text] [doi: 10.2196/12770] [Medline: 31199302]
28. Sathyanarayana A, Joty S, Fernandez-Luque L, Ofli F, Srivastava J, Elmagarmid A, et al. Sleep quality prediction from wearable data using deep learning. JMIR Mhealth Uhealth 2016 Nov 04;4(4):e125 [FREE Full text] [doi: 10.2196/mhealth.6562] [Medline: 27815231]
29. Mitchell SJ, Godoy L, Shabazz K, Horn IB. Internet and mobile technology use among urban African American parents: survey study of a clinical population. J Med Internet Res 2014 Jan 13;16(1):e9 [FREE Full text] [doi: 10.2196/jmir.2673] [Medline: 24418967]
30. Survey System. Sample Size Calculator URL: https://www.surveysystem.com/sscalc.htm [accessed 2020-01-14]
31. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Acad Pathol 2019;6:2374289519873088 [FREE Full text] [doi: 10.1177/2374289519873088] [Medline: 31523704]
32. Buehler L, Rashidi H. Bioinformatics Basics, Application in Biological Science and Medicine, 2nd Edition. Philadelphia: Taylor and Francis Group; 2005.
33. Chen C, Luo J, Parker K. Image segmentation via adaptive K-mean clustering and knowledge-based morphological operations with biomedical applications. IEEE Trans Image Process 1998;7(12):1673-1683. [doi: 10.1109/83.7`30379] [Medline: 18276234]
34. Caruana R, Niculescu-Mizil A. An empirical comparison of supervised learning algorithms. 2006 Presented at: Proceedings of the 23rd International Conference on Machine Learning; 2006; Pittsburgh p. 161-168 URL: https://dl.acm.org/doi/10.1145/1143844.1143865 [doi: 10.1145/1143844.1143865]
35. Guzmán MG, Kourí G. Dengue: an update. Lancet Infect Dis 2002 Jan;2(1):33-42. [doi: 10.1016/s1473-3099(01)00171-2] [Medline: 11892494]
36. Bengio Y. Learning deep architectures for AI. Found Trends Mach Learn 2009;2(1):1-127 [FREE Full text] [doi: 10.1561/2200000006]
37. Chien T, Lin W. Simulation study of activities of daily living functions using online computerized adaptive testing. BMC Med Inform Decis Mak 2016 Oct 10;16(1):130 [FREE Full text] [doi: 10.1186/s12911-016-0370-8] [Medline: 27724939]
38. iHelp. URL: http://www.healthup.org.tw/irs/irsin_e.asp?type1=87 [accessed 2020-01-22]
39. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978 Dec;43(4):561-573. [doi: 10.1007/bf02293814]
40. Lee Y, Lin K, Chien T. Application of a multidimensional computerized adaptive test for a Clinical Dementia Rating Scale through computer-aided techniques. Ann Gen Psychiatry 2019;18:5 [FREE Full text] [doi: 10.1186/s12991-019-0228-4] [Medline: 31131014]
41. Ma S, Wang H, Chien T. A new technique to measure online bullying: online computerized adaptive testing. Ann Gen Psychiatry 2017;16:26 [FREE Full text] [doi: 10.1186/s12991-017-0149-z] [Medline: 28680455]
42. Ma S, Chien T, Wang H, Li Y, Yui M. Applying computerized adaptive testing to the Negative Acts Questionnaire-Revised: Rasch analysis of workplace bullying. J Med Internet Res 2014 Feb 17;16(2):e50 [FREE Full text] [doi: 10.2196/jmir.2819] [Medline: 24534113]

43. Chien T, Lin W. Improving inpatient surveys: web-based computer adaptive testing accessed via mobile phone QR codes. JMIR Med Inform 2016 Mar 02;4(1):e8 [FREE Full text] [doi: 10.2196/medinform.4313] [Medline: 26935793]
44. Hulin C, Drasgow F, Parsons C. Item Response Theory: Applications to Psychological Measurement. Homewood: Dow & Jones Irwin; 1983.
45. Linacre J. An all-purpose person fit statistic? Rasch Measure Transact 1997;11(3):582-583 [FREE Full text]
46. Wu M, Adams R, Wilson M. Acer ConQuest. Melbourne: Australian Council for Educational Research Press; 1998.
47. Saha S. A comprehensive guide to convolutional neural networks—the eli5 way. 2018 Dec 15. URL: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53 [accessed 2020-01-14]

Abbreviations

AUC: area under the curve
BO: burnout
CNN: convolutional neural network
MBI-HSS: Maslach Burnout Inventory–Human Services Survey
MPRSA: matching personal response scheme to adapt for the correct classification
PA: personal accomplishment
ROC: receiver operation characteristic

Edited by C Lovis; submitted 07.10.19; peer-reviewed by S Probst, S Chen; comments to author 15.12.19; revised version received 15.12.19; accepted 31.12.19; published 07.05.20.

Please cite as:
Lee YL, Chou W, Chien TW, Chou PH, Yeh YT, Lee HF
An App Developed for Detecting Nurse Burnouts Using the Convolutional Neural Networks in Microsoft Excel: Population-Based Questionnaire Study
JMIR Med Inform 2020;8(5):e16528
URL: https://medinform.jmir.org/2020/5/e16528
doi:10.2196/16528
PMID:32379050

Original Paper

The Development of a Practical Artificial Intelligence Tool for Diagnosing and Evaluating Autism Spectrum Disorder: Multicenter Study

Tao Chen[1,2], PhD; Ye Chen[3,4], PhD; Mengxue Yuan[1], MS; Mark Gerstein[5,6,7,8], PhD; Tingyu Li[9,10,11,12,13], MM; Huiying Liang[14,15], PhD; Tanya Froehlich[4,16], PhD; Long Lu[1,3,4], PhD

[1]School of Information Management, Wuhan University, Wuhan, China
[2]School of Information Technology, Shangqiu Normal University, Shangqiu, China
[3]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States
[4]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States
[5]Program in Neurodevelopment and Regeneration, Yale University, New Haven, CT, United States
[6]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, United States
[7]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, United States
[8]Department of Computer Science, Yale University, New Haven, CT, United States
[9]Children Nutrition Research Center, Chongqing, China
[10]Children's Hospital of Chongqing Medical University, Chongqing, China
[11]Ministry of Education Key Laboratory of Child Development and Disorders, Chongqing, China
[12]China International Science and Technology Cooperation Base of Child Development and Critical Disorders, Chongqing, China
[13]Chongqing Key Laboratory of Translational Medical Research in Cognitive Development and Learning and Memory Disorders, Chongqing, China
[14]Guangzhou Women and Children's Medical Center, Guangzhou, China
[15]Guangzhou Medical University, Guangzhou, China
[16]Division of Developmental and Behavioral Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States

Corresponding Author:
Long Lu, PhD
School of Information Management
Wuhan University
No 16, Luojiashan Road, Wuchang District
Wuhan, 430072
China
Phone: 86 18986022408
Email: bioinfo@gmail.com

Abstract

Background: Autism spectrum disorder (ASD) is a complex neurodevelopmental disorder with an unknown etiology. Early diagnosis and intervention are key to improving outcomes for patients with ASD. Structural magnetic resonance imaging (sMRI) has been widely used in clinics to facilitate the diagnosis of brain diseases such as brain tumors. However, sMRI is less frequently used to investigate neurological and psychiatric disorders, such as ASD, owing to the subtle, if any, anatomical changes of the brain.

Objective: This study aimed to investigate the possibility of identifying structural patterns in the brain of patients with ASD as potential biomarkers in the diagnosis and evaluation of ASD in clinics.

Methods: We developed a novel 2-level histogram-based morphometry (HBM) classification framework in which an algorithm based on a 3D version of the histogram of oriented gradients (HOG) was used to extract features from sMRI data. We applied this framework to distinguish patients with ASD from healthy controls using 4 datasets from the second edition of the Autism Brain Imaging Data Exchange, including the ETH Zürich (ETH), NYU Langone Medical Center: Sample 1, Oregon Health and Science University, and Stanford University (SU) sites. We used a stratified 10-fold cross-validation method to evaluate the model performance, and we applied the Naive Bayes approach to identify the predictive ASD-related brain regions based on classification contributions of each HOG feature.

Results: On the basis of the 3D HOG feature extraction method, our proposed HBM framework achieved an area under the curve (AUC) of >0.75 in each dataset, with the highest AUC of 0.849 in the ETH site. We compared the 3D HOG algorithm with the original 2D HOG algorithm, which showed an accuracy improvement of >4% in each dataset, with the highest improvement of 14% (6/42) in the SU site. A comparison of the 3D HOG algorithm with the scale-invariant feature transform algorithm showed an AUC improvement of >18% in each dataset. Furthermore, we identified ASD-related brain regions based on the sMRI images. Some of these regions (eg, frontal gyrus, temporal gyrus, cingulate gyrus, postcentral gyrus, precuneus, caudate, and hippocampus) are known to be implicated in ASD in prior neuroimaging literature. We also identified less well-known regions that may play unrecognized roles in ASD and be worth further investigation.

Conclusions: Our research suggested that it is possible to identify neuroimaging biomarkers that can distinguish patients with ASD from healthy controls based on the more cost-effective sMRI images of the brain. We also demonstrated the potential of applying data-driven artificial intelligence technology in the clinical setting of neurological and psychiatric disorders, which usually harbor subtle anatomical changes in the brain that are often invisible to the human eye.

(JMIR Med Inform 2020;8(5):e15767) doi:10.2196/15767

KEYWORDS

autism spectrum disorder; magnetic resonance imaging; neuroimaging; brain; histogram of oriented gradients; cluster analysis; classification; machine learning

Introduction

Background

Autism spectrum disorder (ASD) is a heterogeneous disorder characterized by social impairments, communicative deficits, and restricted, repetitive behaviors. According to the 2018 Centers for Disease Control and Prevention report on autism, approximately 1% (1/59) of US children aged 8 years have been diagnosed with ASD, which represents an increase compared with previous reports [1]. The diagnosis and intervention costs of ASD are growing in concert with the increasing prevalence. A recent study predicted that treatment costs will rise to US $461 billion in 2025 if the prevalence rate of ASD holds steady at present rates and that costs will rise to US $1 trillion by 2025 if the prevalence rate of ASD continues to steeply rise as seen over the last decade [1]. However, concerns have been raised about the accuracy and validity of the reported increase in ASD prevalence, as many other neurobehavioral conditions, as well as variations in developmentally normal behaviors, share common features with ASD and may be misdiagnosed as ASD [2]. Inappropriate ASD diagnoses, and therefore potentially inappropriate applications of ASD-related therapies, stand to increase economic burden. Conversely, deferred or missed ASD diagnosis in children meeting the diagnostic criteria, which appears to be a particular problem for certain sociodemographic [3] and clinical groups [4], lead to a delay in receipt of services and place children at risk for worse outcomes. Therefore, appropriate and early ASD diagnosis and intervention is of crucial importance to improve prognostic outcomes and reduce economic costs.

ASD is now diagnosed mainly by clinical behavior-based approaches, which incorporate standardized tools such as the Autism Diagnostic Observation Scale and Autism Diagnostic Interview-Revised scale. However, this approach is subjective and time consuming [5]. Although it has been reported that ASD has a strong genetic basis, genetic markers are not currently used in the diagnostic process as ASD etiology is complex and the full complement of autism-associated genes is unclear. As magnetic resonance imaging (MRI) is a widely used noninvasive examination method to detect brain abnormalities in clinical practice, there is much interest in its potential to improve or refine the ASD diagnostic process. In clinics, structural MRI (sMRI) has been successfully used to facilitate the diagnosis or treatment of space-occupying lesions such as tumors [6,7]. However, the structural changes of the brain in neurological and psychiatric disorders are not as salient as tumors; thus, it is difficult for clinicians to discover the subtle anatomical changes in the brain. Many studies have focused on finding the functional connectivity abnormalities in the brain using functional MRI (fMRI). Indeed, investigators have explored the use of fMRI to identify ASD. For example, Guo et al [8] developed a deep neural network model using the functional connectivities between brain regions based on the resting-state fMRI. Price et al [9] combined dynamic functional connectivity features in a multinetwork algorithm to classify childhood autism. Huang et al [10] fused multiple functional connectivity networks for ASD diagnosis. However, although fMRI can image cerebral hemodynamics with high spatial resolution, the high cost may limit its potential as a widely used ASD diagnostic tool in clinics [11]. More importantly, it is difficult to interpret the functional connectivity-based results owing to the impact of the underlying brain structure, cognitive state, and subject motion during data acquisition [12]. Furthermore, a recent study suggested that the statistical software used to analyze the raw data from fMRIs might be significantly flawed [13].

Compared with fMRI, sMRI has less data requirements, is more commonly used in clinical settings, and is more amenable to populations for whom compliance is a challenge as it can be performed under sedation. Many ASD sMRI studies have used morphometric features, such as brain surface area, volume, and thickness, to distinguish ASD from control images [14,15]. For example, a recent study of infants at high risk for ASD found hyperexpansion of the cortical surface area and expanded brain volumes in those later diagnosed with ASD [16]. In addition, some studies have made strides toward elucidating ASD brain morphology. Specifically, Bigler et al [17] observed differences in the frontal lobe, parietal lobe, temporal lobe, limbic system,

and cerebellum structures for patients with ASD versus healthy controls.

Related Work

Although sMRI images can provide brain anatomical change information, errors in interpretation can occur owing to difficulty in verifying these subtle changes solely by visual examination. In addition, as there is abundant genetic, phenotypic, and clinical heterogeneity among individuals with ASD, these morphometric features alone are insufficient for diagnosing ASD in clinical settings given that each individual feature is unlikely to be present in the full range of individuals meeting the ASD criteria. To address such barriers, in recent years, machine learning algorithms have been developed to identify underlying brain change patterns in other neurobehavioral conditions marked by similar degrees of heterogeneity. When applying machine learning algorithms to sMRI data, image features representing the sMRI image need to be extracted first. Some of these features are adapted from traditional morphology approaches, while others are developed specifically for machine learning approaches. The traditional morphometric features can be classified into region of interest (ROI), voxel-based morphometry (VBM) [18], surface-based morphometry (SBM) [19], deformation-based morphometry (DBM) [20], and tensor-based morphometry (TBM) [21,22]. Unfortunately, the ROI, VBM, SBM, DBM, and TBM approaches all have significant limitations. Owing to requiring manual or semimanual delineation of brain regions, the ROI process may be labor intensive and time consuming [23]. The performance of VBM, DBM, and TBM methods is highly sensitive to registration accuracy, which is difficult to achieve [24], and is reliant on deformation registration, which may cause over-alignment problems [25]. The SBM method is unable to admit subcortical structures, such as the amygdala and basal ganglia, which may play crucial roles in ASD [26]. To address the limitation of traditional image features discussed earlier, local image features developed specifically for machine learning approaches, such as scale-invariant feature transform (SIFT) [27], do not depend on precise deformation registration. SIFT is assumed to be invariant to image translation, scaling, and rotation and robust to local geometric distortion, which has already been applied to analyze brain images [25,28-31]. However, SIFT itself has several shortcomings. Although SIFT can improve classification accuracy compared with traditional morphometry features, it uses an expert-designed approach to identify visually salient changes that may not relate to the disease. Moreover, SIFT can only describe the characteristics of a limited number of key points and the regions around the key points. However, given that abnormal brain regions in neurodevelopmental disorders/diseases may occur in any position and may be very small, they may be overlooked by the SIFT modality.

Given the above limitations in traditional image features as well as SIFT, another prominent local image feature called histogram of oriented gradients (HOG) [32] has been widely used in computer vision applications (eg, human detection [33,34], vehicle classification [35,36], traffic sign detection [37], pose estimation [38], and general image classification [39]). As HOG can describe the distribution of intensity gradients or edge directions well, it is useful for characterizing local object appearance and shape [32]. In addition, as HOG features can filter most of the nonessential information (eg, a constant colored background) while providing an output of multiple bidimensional histograms for a brain region to reflect the changes within a brain region, HOG features are good at reflecting small or subtle anomalies that may be ignored by SIFT. In prior studies, HOG has generally been used to describe 2D images. Although 2D HOG can be applied to a 3D image, the 3D image needs to be sliced into a series of 2D images along a certain orientation, which can be problematic as changes induced by the disease may be evident only at specific orientations. Fortunately, a recently developed modality called 3D HOG can be analyzed directly inside the 3D volumetric image, which allows image gradient information for the abnormal region to be kept in a more discriminative 3D form and therefore improves classification performance.

Objectives

To address the unique challenges inherent in the neuroimaging studies of ASD, we therefore proposed a novel 2-level classification framework called histogram-based morphometry (HBM), which is based on the 3D HOG feature extraction method. Instead of processing the whole brain image, we divided the entire brain into a few local regions with a given size, which is the foundation of our 2-level hierarchical framework. The first-level classifier is designed for the local regions related to diseased or healthy status, while the second-level classifier or final classifier is for the entire brain that is represented with the concatenation of each region's status. The 3D HOG is computed not for the entire brain but for each local brain region. By using the HBM classification framework, we can classify individuals as patients with ASD or healthy controls. Moreover, the classification contribution of each local HOG feature can be calculated and those features contributing most to the disease classification result can be used to distinguish the predictive brain regions associated with ASD.

This paper has presented the development of the 3D HOG and HBM methods, as well as their application to ASD datasets. In the Methods section, we have described the data source, data preprocessing, 3D HOG feature design, 2-level HBM framework development, and the experimental design. In the Results section, we have discussed the experiment results derived from the analysis of data from the second edition of the Autism Brain Imaging Data Exchange (ABIDE II) [40]. We have concluded by contextualizing our results and discussing the outlook for future ASD neuroimaging research.

Methods

Data Acquisition and Preprocessing

In this study, we used sMRI data from ABIDE II, which includes 19 datasets collected at 18 sites (2 datasets were collected at the same site) and 1114 subjects (521 patients with ASD and 593 healthy controls). For each subject, the ABIDE II datasets consist of resting-state fMRI images, T1-weighted sMRI images, and phenotypic information. Some sites also include diffusion tensor imaging data that may be used to investigate the structural abnormalities of white matter. As an enhancement to the first

edition of the Autism Brain Imaging Data Exchange (ABIDE I) datasets, ABIDE II provides greater phenotypic characterization than ABIDE I data to better address the 2 key sources of heterogeneity: psychiatric co-occurring illness and female sample percentage [40]. The inclusion and diagnostic criteria for patients with ASD and healthy controls are different between each site, and details of the criteria are described in the study by Martino et al [40]. From the 17 datasets, we chose 4 datasets collected from 4 sites, including ETH Zürich (ETH), NYU Langone Medical Center: Sample 1 (NYU), Oregon Health and Science University (OHSU), and Stanford University (SU). Data from a total of 119 patients with ASD and 131 healthy controls from across these 4 sites were used for these analyses. Table 1 lists the sample overview for each site. Age is an important factor that may affect different characteristics, for example, cortical thickness, of the brain in ASD. To evaluate the applicability of our proposed HBM method to different age ranges, we chose the 4 datasets that represent distinct age distributions among all the datasets. Specifically, to reduce the impact of multisite data heterogeneity, we first used single-site data for model classification performance evaluation. Then, we combined all the data from the 4 datasets to evaluate model capability to deal with data heterogeneity.

As the ABIDE II data are original Digital Imaging and Communications in Medicine (DICOM) images, in the first step of data preprocessing, we used the MRIcron tool to convert DICOM images to NifTI images. Then, data processing was performed using SPM12 (UCL Queen Square Institute of Neurology, United Kingdom), which is a third-party package for MATLAB (MathWorks, Natick, Massachusetts, United States). All converted structural images were segmented and normalized to an Montreal Neurological Institute (MNI) standard space.

Table 1. Overview of participants in the 4 training datasets.

Index	Dataset	ASD[a], n (male/female)	Healthy controls, n (male/female)	Age (years), mean (SD)	Age range (years)
1	ETH[b]	13 (13/0)	24 (24/0)	22.7 (4.4)	14-31
2	NYU[c]	48 (43/5)	30 (28/2)	9.8 (4.9)	5.2-34.8
3	OHSU[d]	37 (30/7)	56 (27/29)	10.9 (2.0)	7-15
4	SU[e]	21 (19/2)	21 (19/2)	11.1 (1.2)	8-13
5	Mixed[f]	119 (105/14)	131 (98/33)	12.4 (5.6)	5.2-34.8

[a]ASD: autism spectrum disorder.

[b]ETH: ETH Zürich.

[c]NYU: NYU Langone Medical Center: Sample 1.

[d]OHSU: Oregon Health and Science University.

[e]SU: Stanford University.

[f]Mixed: dataset combining data from all the 4 datasets.

Developing the 3D Histogram of Oriented Gradients Feature

In the process of extending the concept of HOG from a 2D space to 3D space, we needed to define the methods for calculating the image gradient (including direction and magnitude) and partitioning the gradient directions into a few orientation bins (or channels) in a 3D space. The gradient directions in the 3D space were represented by using 2 angles, theta and phi, as shown in Figure 1. Then, the gradient of each image voxel is calculated based on these 2 angles (see Multimedia Appendix 1 for more details).

Similar to 2D HOG, the gradient direction in 3D HOG also needed to be partitioned into several orientation bins. The difference lies in that the partitions in 2D HOG are spread over 360° in just one 2D plane, while the partitions in 3D HOG are spread over the entire volumetric space. There are many partition schemes to divide the orientation space. We have introduced the 2 partition schemes as follows.

The first scheme is to allocate the orientation bins in horizontal and vertical directions with equal-space angle ranges, such as the 2D HOG, and each bounded area between the 2 directions is considered as one 3D partition. The partition results are shown in Figure 2.

When every partition area is projected onto the sphere surface, they correspond to the surface area between the latitude and longitude lines. For this partition scheme, the number of orientation bins, which is equal to the dimension number of the 3D HOG features, is calculated using the following equation in:

where N_{DIR3} is the number of directions in 3D space and N_{DIR2} is the number of directions in 2D space.

In Figure 2, part (a), for the partitions near the poles, a slight change in the angles will result in a different orientation bin assignment. This causes the features to be overly sensitive to the angle differences in some but not all directions. To avoid potential performance loss because of this phenomenon, we proposed an additional partition scheme, in which the partitions adjacent to the pole points are combined into 1 partition as shown in Figure 2, part (b).

The number of orientation bins for this second partition scheme, which merges the direction areas near the pole into 1 direction, is calculated using the equation in:

For the convenience of calculation, the value of N_{DIR2} is constrained to be an even number. For example, if N_{DIR2} is set to 8, N_{DIR3} will be 32 as calculated in the first scheme while in the second scheme N_{DIR3} will be 26.

Figure 1. Two angles related to gradient direction calculation in 3D space.

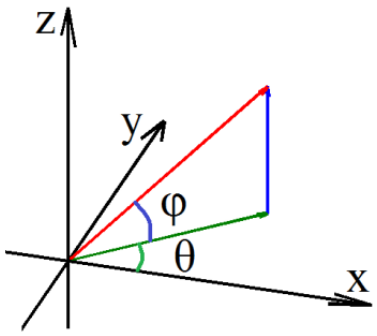
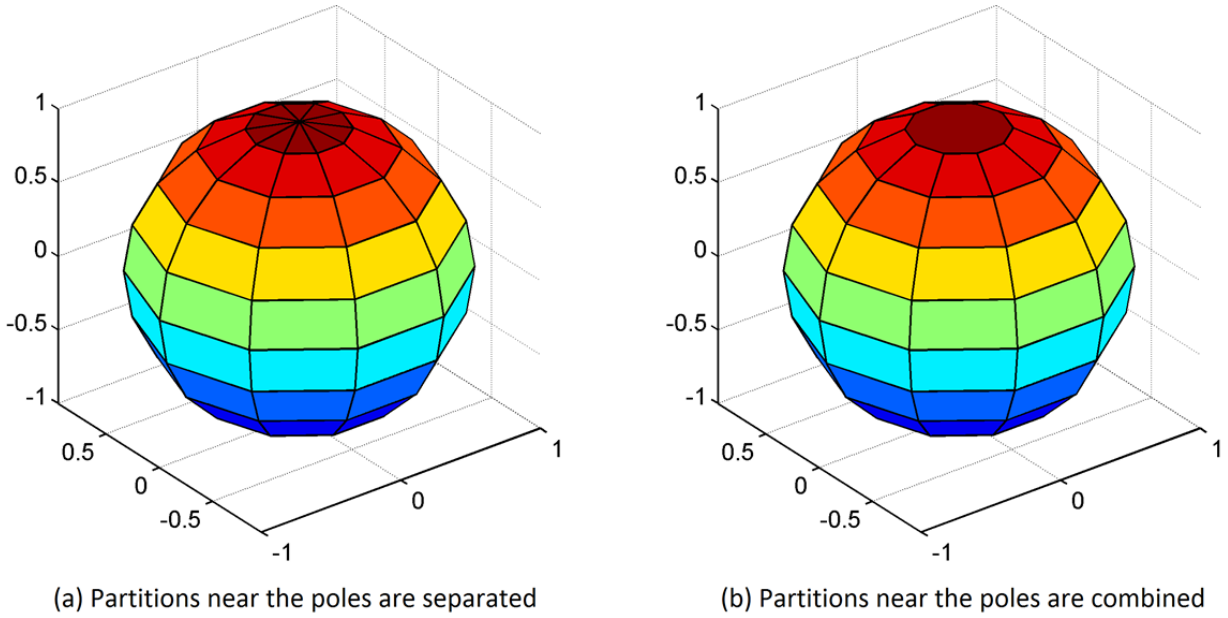


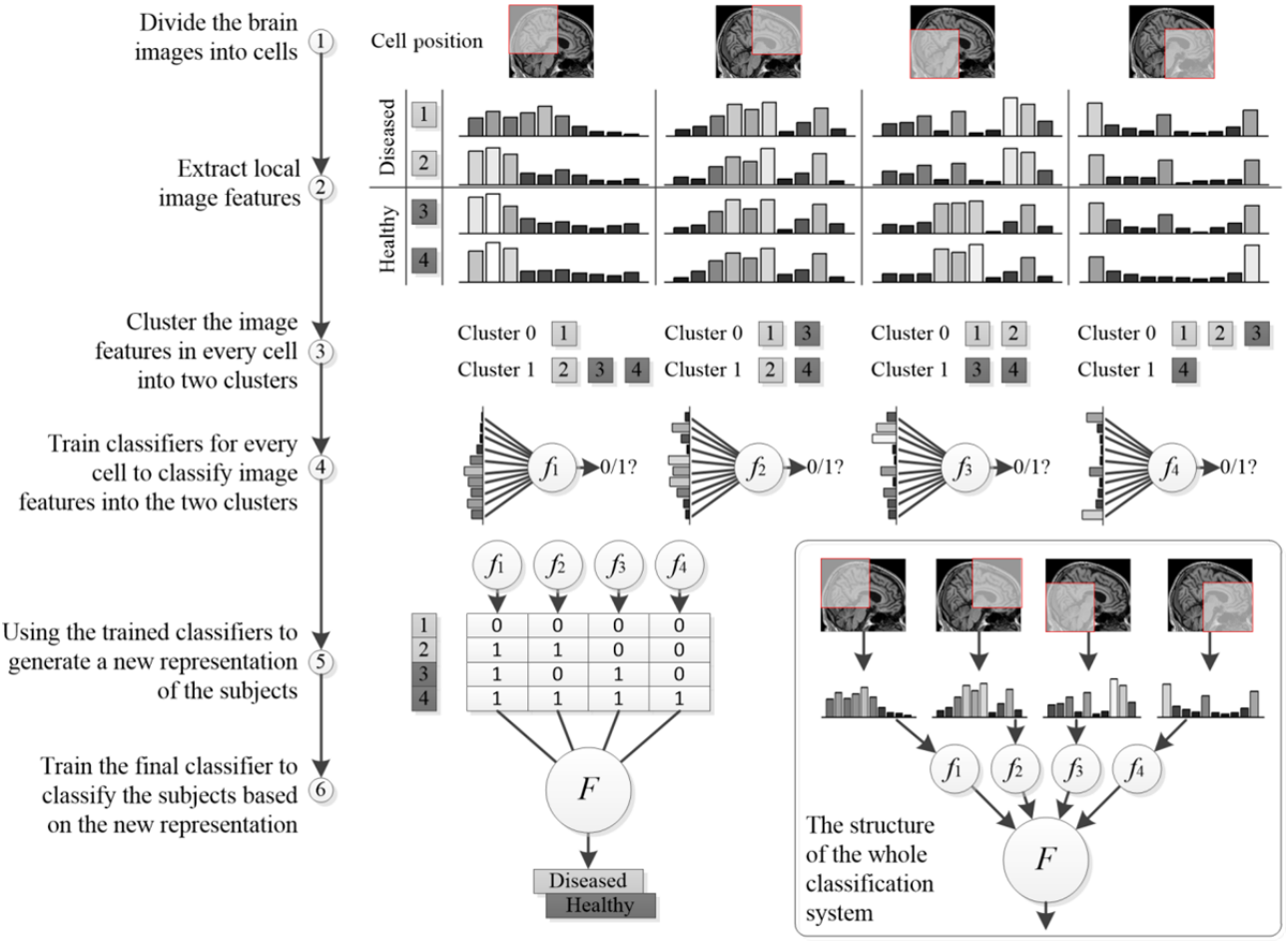
Figure 2. Two partition schemes of the orientation bins in 3D space.



Overall Classification Framework

In this paper, we proposed a 2-level HBM classification framework based on 3D HOG features to differentiate between patients with ASD and healthy controls. Each brain image was firstly divided into a densely overlapping grid of regional cells, and the 3D HOG feature of each cell was computed. On the basis of the brain division, we developed a first-level classification algorithm to predict whether a given cell provides strong evidence to support a final disease/health classification. As there is no label for each cell, a clustering algorithm was used to first find the labels for each cell (the details have been discussed in the following sections). Then, a second-level classification was used to make a final classification based on all the evidence from each cell. Figure 3 shows the 2-level classification framework using a 2D image example for convenient illustration. The bottom-right part of the figure represents the testing process, while the remaining part shows the training process.

Figure 3. Overview of the proposed histogram-based morphometry (HBM) classification framework.



Algorithm Steps

Brain Image Division and Local Feature Extraction

Before the feature extraction step, we first divided the entire 3D MRI brain image into regional cells in step 1. This brain division method can be applied not only to 3D MRI volumetric images, in which a regional cell equates to a *cube*, but to 2D MRI slices, in which a regional cell equates to a *square*. In our algorithm, we computed the HOG feature for each cell but did not collect it into a combined feature vector used to represent the entire image. In the standard HOG usage, all the local HOG features were combined into a high-dimensional feature vector used as input to the classifier [32]. In our hierarchical classification framework, these local features were transformed into high-level forms that can reduce the dimensionality of the features input to the final classifier, which has the benefit of reducing overfitting in the relatively small-sized datasets that are often available in medical studies. Furthermore, using local features is helpful to identify the ASD-related brain regions that have large feature contributions to the disease classification result. In image division, cell size and cell overlapping percentage are 2 important parameters that will affect the classification accuracy. Therefore, different brain image division schemes should be evaluated to determine which has the best classification performance.

In step 2, we extracted local HOG features using 2 different gradient direction partition schemes: HOG-32 and HOG-26, as shown in parts (a) and (b) in Figure 2, respectively. A comparison between these 2 schemes is also necessary to determine which has superior performance. Of note, better classification performance using the 3D HOG algorithm usually results from MRI scans with high spatial resolution, while the performance of the 3D HOG algorithm may degrade if the MRI scan has a low spatial resolution. In this case, an alternative 2D HOG algorithm may be used.

Local Feature Clustering and Regional Classifier Training

In step 3, we worked on each cell independently. For each cell, the goal was to find a binary representation to indicate whether it is related to the diseased status or healthy status. However, we did not have a class label for each cell. Although the class label of the whole brain is known in training samples, it does not mean that each cell should have the same class label as the whole brain. Even in a diseased subject, there may be a lot of cells in the brain that look perfectly normal. Owing to the unknown class label for each cell, we applied a clustering algorithm to the training samples to get the class labels of individual cells. As the distribution of clusters is unknown, we tried 2 different clustering algorithms, such as K-means and hierarchical clustering, that are suitable for different cluster distributions. Although the clustering algorithm works well during the training stage, we proposed to use a classification algorithm to generate the binary representation during the testing stage. The reason we used classification instead of clustering

during the testing stage was because we did not need to keep all the training features while using the approach to make a prediction, which makes the method more scalable and practical. Thus, based on the clustering labels of cells in training samples, we built regional classifiers in step 4 for predicting the cell status of test samples. When the K-means algorithm is used for clustering, the resulting clusters usually have a spherical shape in feature space and the centroids are good exemplars for the corresponding clusters. Therefore, the nearest centroid classification method was used in this case. If the hierarchical algorithm is used for clustering, the centroids of the clusters may not be representative of the cluster, and therefore, the nearest centroid classifier is not appropriate. In this case, the support vector machine (SVM) can be used to build regional classifiers for testing samples.

Compact Feature Representation and Final Classifier Training

The labeled local features only reflect the status of brain regions and not the whole picture of the characteristics of the brain. Therefore, in step 5, we concatenated each local feature status of 1 brain image into a new high-level compact feature representation of that image. For model training, we constructed the high-level feature by directly concatenating the clustering results obtained in step 4. Of note, the clustering result of each feature was concatenated according to a certain sequence, for example, from top-left to bottom-right on the grid. Such a sequence is actually determined by the HOG feature extraction algorithm, and the same sequence is also used when concatenating the binary status of HOG features, thus ensuring the unified meaning of feature representation for all samples. On the basis of the new feature representation and diagnosis labels of the training data, we trained the final classifier using the SVM classification method in step 6. SVM is one of the most widely used classifiers that can perform not only linear classification but also nonlinear classification [41]. It has already been applied to various diseases and neurodevelopmental disorders, for example, Parkinson disease [42], Alzheimer disease [43,44], ASD [45,46], attention-deficit/hyperactivity disorder [47], and schizophrenia [48].

Process for the Test Sample Classification

The abovementioned steps describe the whole training process of obtaining the 2-level classification models including the regional classifier and the final classifier. We could then apply these classifiers to unknown test samples. First, 3D local HOG features of the cells in a test brain image are extracted with the same method as the training process. Then the regional classifiers, such as the nearest centroid, are used to classify each local HOG feature into disease-related or healthy-related labels. These labels are then concatenated to generate the compact representation of that test image. Finally, the final classifier is applied to predict whether the test sample is a patient with ASD using the compact feature vector as the input to the classification model.

Feature Contribution Calculation

Besides using the HBM framework to make a classification of the test sample, we could also investigate each cell's feature contribution to the algorithm's prediction that each participant is a patient with ASD versus a healthy control. A higher value of the feature contribution indicates more likelihood of a cell being disease-related. As we used the SVM method in the final classification level, the feature contribution could be calculated based on the coefficients of the linear SVM classifier. However, this method can cause problems as we do not know which clustered label represents the diseased status. Thus, we chose the Naive Bayes approach instead to calculate the feature contribution for both clustered labels. In the strictest sense, the *feature contribution* calculated by the Naive Bayes method should be called *feature importance*, which only reflects the feature contribution given that the final classifier is a Naive Bayes classifier. We will explore more interpretable mapping from the local features to the final classification results in future research.

First, we will introduce the Naive Bayes approach, which is based on Bayes' theorem. This approach has been widely used for classification in many domains owing to its simplicity and strong performance. It is assumed that predictive features X_0, X_1, …, X_n are independent of each other given the state of a class variable Y. Although it is difficult to reduce the dependence for a neuroimage analysis because different brain regions are correlated in many ways by nature, empirical observations have suggested that the Naive Bayes works quite well even when there is dependence between features [49]. Therefore, we used Bayes' theorem to derive the posterior probability $P(Y \mid X_0, X_1, \ldots, X_n)$ as follows:

where $X_i \in \{0, 1\}$ represents the ith cell clustering result, and $Y \in \{D, H\}$ represents the training sample label. In addition, P_D and P_H refer to the probability of being classified as a patient with ASD versus a healthy control, respectively, conditioned on the state of each cell. If $P_D > P_H$, we predicted that the test sample is more likely to be a patient with ASD than a healthy control. To avoid underflow in the Bayesian computation, we used the log ratio as follows:

where we defined the log sum item  as the feature contribution at the ith cell. A higher value of this item indicates a more predictive feature. It is worth noting that because we did not know exactly which cell state (0 or 1) indicates a disease-related feature and these 2 feature states can both contribute to the disease, we calculated both of their feature contributions.

Then, according to the first-level classification results of each cell in a test patient sample, the most predictive features whose contribution values are above a preset threshold can be identified. We set a threshold on the feature contribution to just show the top features to the patients (in a hypothetic clinical use case). The threshold is usually set to different values when

using heterogeneous sMRI data from different sites or when the parameter values (eg, cell size and cell overlapping percentage) are changed. The cells that contribute most to the classification result of ASD are considered to be the candidate regions related to the disease.

Experimental Design

In the 2-level HBM framework, we evaluated the 2 different 3D gradient direction partition schemes using the algorithm combinations for feature clustering, regional classifier training, and final classifier training listed in Table 2. The performance of the 4 instances listed in the table will be compared later. The instance name in the table (eg, KNS32) is the abbreviation created using the first letter from the local feature clustering algorithm name (K-means), the regional classification algorithm name (nearest centroid), the final classification algorithm name (SVM), and 32 orientation bins.

After the final classification model is trained, its performance is evaluated, typically via the cross-validation (CV) method. The widely used CV methods in brain image analysis include leave-1-out CV [25,48,50], leave-2-out CV [45,51,52], k-fold CV [53,54], and stratified k-fold CV [55,56]. Although there are conflicting reports in the literature, most papers, including a review of brain image classification methods, suggest that 10-fold CV is the most appropriate method [57]. In this study, we trained our model using the stratified 10-fold CV method. The stratified CV method provides the following advantages. First, the stratified method can keep the ratio of 2 sample classes in each fold as close to that of all samples as possible, retaining the original data distribution pattern of the entire dataset. Second, the variance of model performance estimations will decrease by performing several random runs, in each of which all samples are first shuffled and then split into a pair of training and test sets. The stratified CV method proposed in this paper is implemented as the pseudo-code shown in Figure 4.

In the 3D HOG partition scheme, there is a parameter N_{DIR2} that represents the number of orientation bins in either the horizontal or vertical direction of the 3D space. If N_{DIR2} is set too high, the computation speed of the algorithm will be slowed. However, more importantly, the feature will be more sensitive to noise and other noninformative signals in the images. Furthermore, the dimension of the feature will be high, which usually requires more samples to avoid the *curse of dimensionality*. Otherwise, if N_{DIR2} is set too low, details of the image will be lost. In this paper, we set the number of N_{DIR2} to the frequently used value 8, and the total number of directions in 3D space was 32 and 26 for the two 3D HOG partition schemes. The other parameters for the HOG features, including cell size and overlapping percentage, were evaluated using the CV method. The performance measures we used to evaluate our algorithm included classification accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1 score, and the area under the curve (AUC).

Table 2. The 4 instances of the proposed histogram-based morphometry framework used for performance evaluation.

Instance name	Image feature	Image feature processing for each cell		Final classification
		Clustering	Classification	
KNS32	HOG[a]-32[b]	K-means	Nearest centroid	SVM[c,d]
KNS26	HOG-26[e]	K-means	Nearest centroid	SVM[c]
HSS32	HOG-32[b]	Hierarchical	Linear kernel SVM	SVM[c]
HSS26	HOG-26[e]	Hierarchical	Linear kernel SVM	SVM[c]

[a]HOG: histogram of oriented gradients.

[b]HOG-32 is the histogram of oriented gradients feature with 8 directions in a 2D plane and 32 directions in 3D space.

[c]Three different kernels have been tested, for example, the linear kernel, the polynomial kernel, and radial base function kernel.

[d]SVM: support vector machine.

[e]HOG-26 is the HOG feature with 8 directions in a 2D plane, and the 2 poles are considered as 2 directions in 3D space; therefore, the total number of directions is 26.

Figure 4. Algorithm of the stratified cross-validation with multiple random runs.

Input: number of folds N_F, number of cross-validations N_{CV}.
Output: mean μ and standard deviation σ of the classification accuracies.
Steps:
Calculate the ratio of the two classes R_C;
for $cur\_cv = 1$ to N_{CV} **do**
Partition the samples into N_F folds, with the ratio of two classes in each fold as close to R_C as possible;
for $k = 1$ to N_F **do**
Train the classifier on the samples that are not in fold k;
Test the classifier on the samples in fold k;
end for
Calculate the classification accuracy in current cross-validation R_{cur_cv};
end for
Calculate the mean and standard deviation of the classification accuracies for all cross-validation experiments, i.e., $\mu = \frac{1}{N_{CV}} \sum_{n=1}^{N_{CV}} R_n$, and $\sigma = \sqrt{\frac{1}{N_{CV}} \sum_{n=1}^{N_{CV}} (R_n - \mu)^2}$.

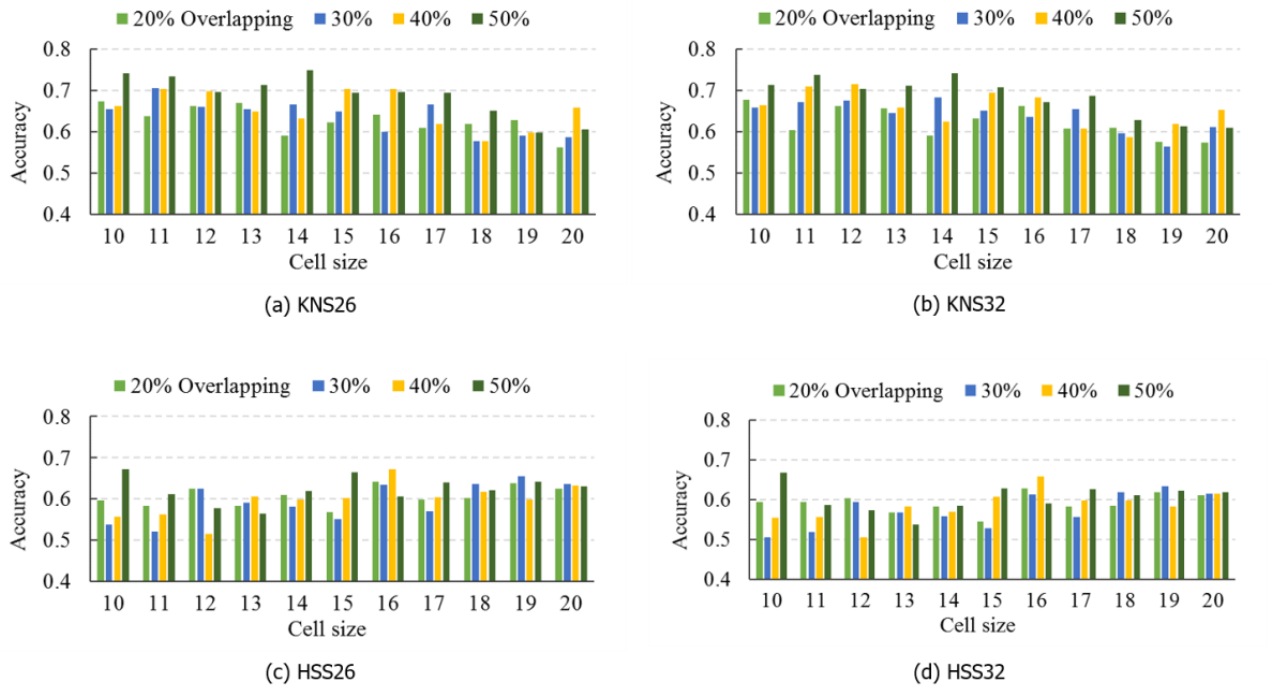
Results

Comparing the Classification Performance of Different Histogram-Based Morphometry Instances

To compare the performance of the 4 HBM instances listed in Table 2, we used the stratified 10-fold CV evaluation method to obtain each performance measure. As the size of cell and the overlapping between 2 cells may influence the model's performance, we performed a parameter scan for the best values of these 2 parameters. The cell size ranged from 10 voxels to 20 voxels and cell overlapping percentage ranged from 20% to 50%. In the final classification step, we tested 3 different SVM kernels, including the linear kernel, the polynomial kernel, and radial base function kernel. We then chose the linear kernel for use owing to its superior performance.

Figure 5 shows the stratified 10-fold CV average accuracies based on the data from the NYU site when using different HBM instances and different parameter values. The expanded form of the abbreviations of the HBM instances in Figure 5 can be found in Table 2. From the figure, it can be seen that although the classification accuracies fluctuate as the parameter values change, KNS26 and KNS32 performed significantly better than HSS26 and HSS32, which means that the combination of K-means and centroid algorithms is more suitable for our proposed HBM framework. Meanwhile, Figure 5 shows that KNS26 outperformed KNS32 and HSS26 outperformed HSS32, which supports the rationality and effectiveness of the HOG-26 partition scheme. In addition, among the different parameter values, KNS26 obtained the best average classification accuracy, 74% (58/78), when the cell size was set to 14 voxels and the cell overlapping percentage was set to 50%. For the other 3 sites, ETH, OHSU, and SU, KNS26 also outperformed KNS32, although the best parameter values may be different (see Multimedia Appendix 2 for the results of these additional analyses). Of note, our method was not overly sensitive to the parameters, so model performance was generally good for a wide range of parameters.

Figure 5. Classification accuracies for the NYU Langone Medical Center: Sample 1 dataset using 4 histogram-based morphometry (HBM) instances including KNS26 (a), KNS32 (b), HSS26 (c), and HSS32 (d).



Comparing the Classification Performance of Different Local Feature Extraction Algorithms

In this paper, we used the HOG algorithm for local image feature extraction in the HBM framework. This algorithm helps to generate high-quality representations that depict image edge and texture. To evaluate the effects of different local feature extraction algorithms on classification performance, we also used SIFT, another widely used local feature detection algorithm, to extract features from brain images and developed an SVM approach to analyze the extracted SIFT features. This approach has been applied to neurological diseases such as Alzheimer disease [25,31], Parkinson disease [31], and bipolar disease [31]. As shown in Figure 5, KNS26 was the best performing HBM instance, so we compared it (rather than KNS32) with the SIFT-based SVM approach.

We trained both classifiers using the stratified 10-fold CV, and the training data were the same for them in each fold. The results show that a HOG-based KNS26 HBM approach achieves much better performance than the SIFT-based SVM approach (Tables 3 and 4). Overall, comparison results depicted in Tables 3 and 4 demonstrate that HOG features are more suitable for delineation of the underlying structural change patterns in sMRI images than SIFT features. By transforming the low-level HOG features into high-level features, our proposed 2-level HBM classification framework can effectively employ the high-level features to differentiate individuals as either patients with ASD or healthy controls. In the last row of Table 3, we can see that the performance degraded when building the model on data from the 4 datasets. We have discussed the reason in the Discussion section.

Table 3. Classification performance using histogram-based morphometry on the second edition of the Autism Brain Imaging Data Exchange datasets.

Dataset	Best parameter		Histogram-based morphometry (KNS26)											
	Cell size	Overlapping (%)	ACC[a]		SEN[b]		SPE[c]		PPV[d]		NPV[e]		F1[f]	AUC[g]
			N	n (%)	N	n (%)	N	n (%)	N	n (%)	N	n (%)		
ETH[h]	10	20	37	32 (86)	13	10 (77)	24	22 (92)	12	10 (83)	25	22 (88)	0.790	0.849
NYU[i]	14	50	78	58 (74)	48	40 (83)	30	18 (60)	52	40 (77)	26	18 (69)	0.805	0.787
OHSU[j]	19	40	93	70 (75)	37	23 (62)	56	46 (82)	33	23 (70)	60	46 (77)	0.662	0.794
SU[k]	17	20	42	30 (71)	21	17 (81)	21	13 (62)	25	17 (68)	17	13 (77)	0.751	0.763
Mixed[l]	12	30	250	162 (65)	119	87 (73)	131	76 (58)	142	87 (61)	108	76 (70)	0.662	0.650

[a]ACC: accuracy is the ratio of correctly classified subjects over all subjects.

[b]SEN: sensitivity is the ratio of correctly classified subjects with autism spectrum disorder (ASD) over all subjects with ASD.

[c]SPE: specificity is the ratio of correctly classified subjects without ASD over all subjects without ASD.

[d]PPV: positive predictive value is the ratio of correctly classified subjects with ASD over all predicted subjects with ASD.

[e]NPV: negative predictive value is the ratio of correctly classified subjects without ASD over all predicted subjects without ASD.

[f]F1: F1 score.

[g]AUC: area under the curve.

[h]ETH: ETH Zürich.

[i]NYU: NYU Langone Medical Center: Sample 1.

[j]OHSU: Oregon Health and Science University.

[k]SU: Stanford University.

[l]Mixed: dataset combining data from all the 4 datasets.

Table 4. Classification performance using scale-invariant feature transform and support vector machine on the second edition of the Autism Brain Imaging Data Exchange datasets.

Dataset	Performance using scale-invariant feature transform and support vector machine											
	ACC[a]		SEN[b]		SPE[c]		PPV[d]		NPV[e]		F1[f]	AUC[g]
	N	n (%)	N	n (%)	N	n (%)	N	n (%)	N	n (%)		
ETH[h]	37	24 (65)	13	8 (62%)	24	16 (67)	16	8 (50)	21	16 (76)	0.533	0.709
NYU[i]	78	44 (56)	48	29 (60)	30	15 (50)	44	29 (66)	34	15 (44)	0.624	0.595
OHSU[j]	93	52 (56)	37	19 (51)	56	33 (59)	42	19 (45)	51	33 (65)	0.482	0.605
SU[k]	42	18 (43)	21	10 (48)	21	8 (38)	23	10 (44)	19	8 (42)	0.449	0.367

[a]ACC: accuracy is the ratio of correctly classified subjects over all subjects.

[b]SEN: sensitivity is the ratio of correctly classified subjects with autism spectrum disorder (ASD) over all subjects with ASD.

[c]SPE: specificity is the ratio of correctly classified subjects without ASD over all subjects without ASD.

[d]PPV: positive predictive value is the ratio of correctly classified subjects with ASD over all predicted subjects with ASD.

[e]NPV: negative predictive value is the ratio of correctly classified subjects without ASD over all predicted subjects without ASD.

[f]F1: F1 score.

[g]AUC: area under the curve.

[h]ETH: ETH Zürich.

[i]NYU: NYU Langone Medical Center: Sample 1.

[j]OHSU: Oregon Health and Science University.

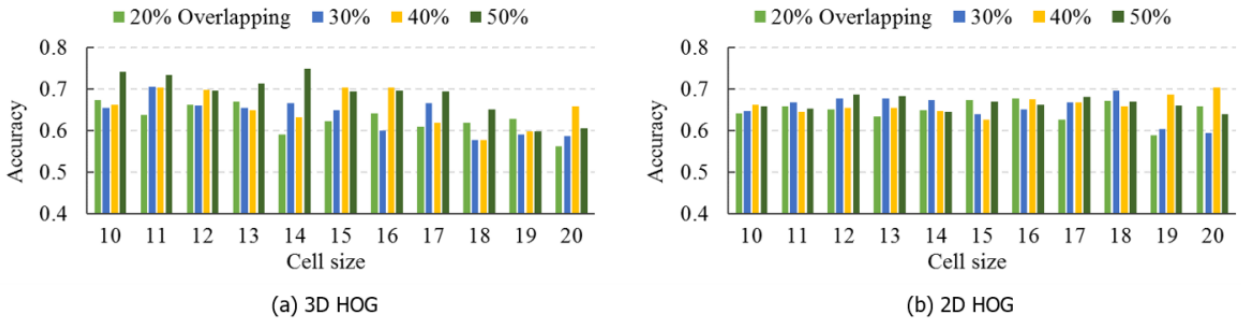
[k]SU: Stanford University.

Comparing 3D Histogram of Oriented Gradients and 2D Histogram of Oriented Gradients

HOG features represent image edge and texture, and the feature quality is affected by MRI acquisition parameters, especially spatial resolution that is decided by slice thickness, matrix size, and field of view. Low spatial resolution will decrease image sharpness and cause fuzzy edges, which may degrade the classification performance. By contrast, high spatial resolution helps to retain more fine-grained and high-contrast information of the brain tissues, which enable us to extract HOG features directly in its inherent 3D form. From the anatomical scan parameters, we can see that the T1-weighted sMRI images are all high-resolution images in these 4 datasets. In our proposed 3D HOG algorithm, the features were extracted directly inside the 3D volumetric image. In the 2D HOG algorithm, the features were extracted from the 2D MRI slices. The hypothesis is that the 3D HOG algorithm will generate highly discriminative representations with higher quality than those generated by the 2D HOG algorithm.

To validate the hypothesis, we tested all the HBM instances listed in Table 2 for the 4 datasets. Here, data from the NYU site and KNS26 instance are used as examples to compare 3D HOG with 2D HOG. The evaluation scheme for both algorithms was the 10-fold CV, and the same parameter scan scope was used as discussed in the Comparing the Classification Performance of Different Histogram-Based Morphometry Instances section. Figure 6 presents the classification accuracy obtained from 3D HOG and 2D HOG. We can see from the figures that 3D HOG outperforms 2D HOG for some scan parameters and obtains the highest accuracy when the cell size is set at 14 voxels and cell overlapping percentage is set at 50%. The other 3 sites show a comparison result similar to NYU (see Multimedia Appendix 2 for the results of these additional analyses). Thus, the comparison between these 2 HOG algorithms supports the hypothesis that 3D HOG can generate more competitive representations for the ASD diagnosis task.

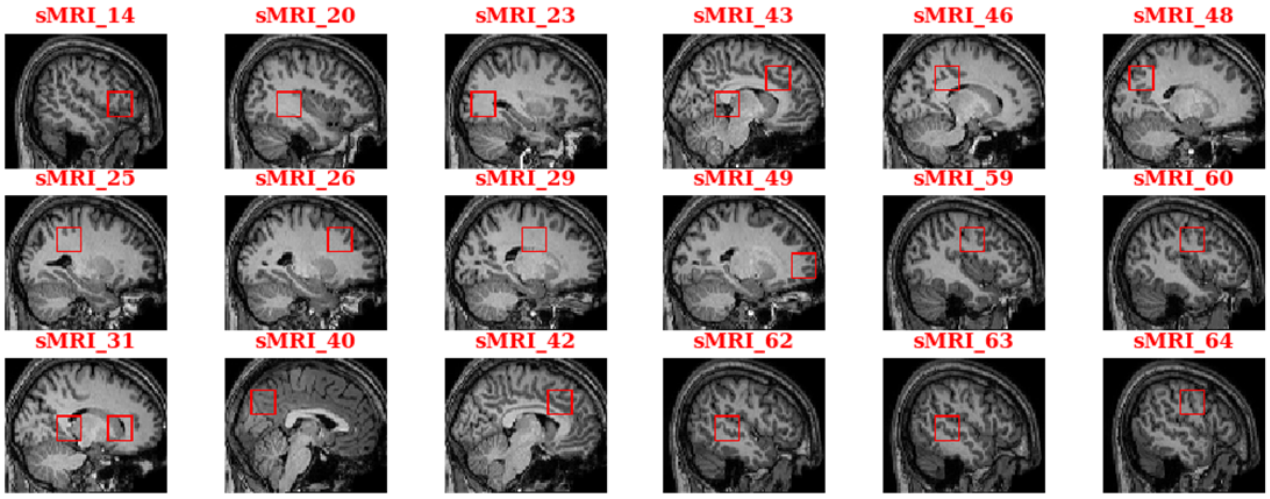
Figure 6. Classification accuracies for the NYU Langone Medical Center: Sample 1 dataset using a 3D histogram of oriented gradients (HOG; a) and 2D HOG (b).



Identifying Predictive Autism Spectrum Disorder–Related Brain Regions

Those predictive features contributing most to the classification prediction of being a patient with ASD versus a healthy control were identified by calculating each cell's feature contribution. Then, the abnormal regions identified as algorithm *high contribution features* were annotated automatically on the MRI image according to the cell's voxel-based coordinates. Figure 7 shows the annotation of the abnormal regions of 1 specific patient with ASD from the ETH dataset. For the convenience of illustration, we annotated these regions in the form of 2D slices. In Figure 7, the number suffix of the legend on top of each slice is the slice number, and each rectangle with the red border indicates an ASD-related region.

Figure 7. Annotation of the autism spectrum disorder–related brain regions for a sample in the ETH dataset. sMRI: structural magnetic resonance imaging.



To give a sound biological interpretation of our results, we located the standard brain regions defined in the anatomical automatic labeling (AAL) brain atlas, which is one of the most widely used cortical parcellation maps. As the AAL brain atlas is constructed on an MNI-based coordinate system, we transformed the coordinates from the voxel space into the MNI space using an affine transformation. Table 5 lists the union of ASD-related regions for all patients in the ETH dataset. The table columns *X*, *Y*, and *Z* represent the central coordinates of the disease-related cells in a 3D MNI-based space. The brain region names in the table are located based on the central coordinates. Owing to the unique set of sulcal folds for each individual, we assigned the closest region to the cell if the cell's center did not fall in any AAL atlas region. The same method can be applied to the other 3 datasets to identify the ASD-related brain regions relevant to each dataset, and the findings show the consistency between these datasets.

Table 5. Autism spectrum disorder–related anatomical automatic labeling brain regions identified by a histogram-based morphometry framework on the ETH dataset.

Index	Region name	Central Montreal Neurological Institute–based coordinates[a]			Studies	
		X	*Y*	*Z*	Guo et al [8]	Huang et al [10]
1	Frontal_Inf_Tri_R	50	22	4	Y	N
2	Temporal_Sup_R	38	−38	4	N	N
3	Calcarine_R	32	−68	4	N	Y
4	Postcentral_R	28	−38	34	N	Y
5	Frontal_Mid_R	26	22	34	Y	Y
6	Caudate_R	20	−8	34	N	N
7	Precuneus_R	16	−38	4	N	Y
8	Caudate_R	16	22	4	N	N
9	Precuneus_L	−2	−68	34	N	Y
10	Cingulum_Mid_R	−6	22	34	Y	N
11	Precuneus_L	−8	−38	4	N	Y
12	Cingulum_Mid_L	−8	22	34	Y	N
13	Cingulum_Mid_L	−14	−38	34	Y	N
14	Precuneus_L	−18	−68	34	N	Y
15	Frontal_Sup_L	−20	52	4	Y	Y
16	Postcentral_L	−42	−8	34	N	Y
17	Temporal_Mid_L	−48	−38	4	N	N
18	Postcentral_L	−50	−8	34	N	Y
19	Lingual_R	18	−68	4	N	N
20	Insula_R	46	−8	4	Y	Y
21	Cingulum_Ant_L	−2	52	4	Y	Y
22	Pallidum_R	26	−8	4	N	N
23	Frontal_Sup_Medial_R	8	52	4	Y	Y
24	Occipital_Mid_R	−32	−68	34	N	Y
25	Parietal_Inf_L	−36	−38	34	N	N
26	Temporal_Sup_L	−50	−8	4	N	N
27	Lingual_L	−12	−68	4	N	N
28	Hippocampus_L	−24	−38	4	N	N
29	Temporal_Mid_R	−46	−38	4	N	N
30	Hippocampus_R	28	−38	4	N	N
31	Cingulum_Ant_R	16	22	34	Y	Y

[a]*X*, *Y*, and *Z* represent the central Montreal Neurological Institute–based coordinates of each disease-related cell that is located in the closest anatomical automatic labeling region. The last 2 columns represent the overlapping brain regions between our study and 2 functional magnetic resonance imaging (fMRI)–based studies (*Y* means a brain region overlaps with the fMRI-based study, whereas *N* means the opposite).

Discussion

Principal Findings

In this study, we developed an innovative 2-level HBM classification framework for distinguishing patients with ASD from healthy controls based on sMRI data and the 3D HOG feature extraction method. Of note, many of the brain regions utilized in our algorithm to indicate ASD—such as frontal gyrus, temporal gyrus, cingulate gyrus, postcentral gyrus, precuneus, caudate, and hippocampus—have been implicated in autism in prior neuroimaging literature [8,58-63]. Currently, ASD is a behaviorally defined disorder, diagnosed through careful clinical assessment. Our intention is not to replace the diagnostic criteria but to begin developing more objective tools which may someday augment the current ASD diagnostic process. At this

juncture, we provide a proof of principle that it may be possible to develop an ASD computer-aided tool based on sMRI images alone by utilizing machine learning techniques. Of note, these techniques offer novel ways to examine neuroimaging data to probe additional clues regarding the neural underpinnings of the disorder.

Although machine learning techniques have been used in prior ASD neuroimaging studies, it is striking that most of these previous studies used fMRI rather than sMRI approaches. Our sMRI approach may represent a significant advancement given that the high cost and lower availability of fMRI likely limits its clinical applicability, while developing clinical approaches to ASD diagnosis that incorporate sMRI may be more practical given sMRI's smaller data requirements, lower cost, and higher clinical availability. Furthermore, given that fMRI evaluates brain activation by measuring cerebral blood flow, typically during the completion of informative tasks, it is often not amenable to use for individuals with ASD. Patients being evaluated for ASD are particularly likely to have difficulty adhering to directions to complete tasks and remain still during fMRI given that they are usually children and have cognitive and/or behavioral impairments that have prompted the diagnostic evaluation. On the contrary, these concerns are well-addressed by the well-developed sedation protocols available for sMRI. In this project, using the more cost-effective sMRI approach, our ASD classification results (32/37, 86% accuracy for the ETH site) were comparable to more expensive and cumbersome fMRI approaches. For example, 2 fMRI studies based on the ABIDE I datasets have been conducted: Huang et al [10] achieved an ASD classification accuracy of 79%, while the fMRI study from Guo et al [8] obtained a classification accuracy of 86%. It should be noted that these 2 studies also used 1 site.

Of note, using our sMRI approach, we identified ASD-related brain regions that overlap with brain regions pinpointed in the above 2 fMRI studies. For example, Guo et al [8] detected ASD-associated brain function connectivities in regions, such as the inferior and superior frontal cortex, temporal cortex, cingulate cortex, and insula, which were also found to be associated with ASD in our study. Similar to Huang et al [10], we also implicated the middle frontal gyrus, middle occipital gyrus, superior frontal gyrus, calcarine cortex, and insula in ASD. The last 2 columns of Table 5 show the overlapping brain regions between our method and the above 2 fMRI-based studies. In the table cell, *Y* means a brain area identified by our method that is also reported in the studies by Guo et al [8] and Huang et al [10] and *N* means the opposite. These brain regions found to be associated with ASD by our study have striking functional correlates with the autism spectrum phenotype. Specifically, regions such as the superior temporal cortex, inferior frontal cortex, several regions of the cingulum, and the insula have been linked to social cognition and language [64]. Variations in the superior temporal gyrus have been linked to ASD-related deficits in the theory of mind (the ability to attribute mental states, such as desires and beliefs, to the self and others [65]) and face processing [66]. The inferior frontal gyrus has been associated with social functioning (including processing of facial expressions [67]) and language processing [68]. The anterior cingulate cortex has been implicated in ASD-related social impairment and repetitive behaviors [68], while the insula is involved in affective and empathic processes [69].

Strengths and Limitations

In addition, our work represents advances over previous sMRI-based ASD neuroimaging studies, as those approaches have typically been limited by the extracted morphometry measures, such as cortical surface area and cortical thickness [16]. Importantly, these sMRI approaches are often unable to probe subcortical features, such as the amygdala and basal ganglia, which have demonstrated importance in ASD and other brain-based disorders such as Parkinson disease and depression. Our approach is amenable to the full breadth of brain structures implicated in ASD and can be easily adapted for use in other brain-based disorders. Indeed, the sMRI-based machine learning algorithm methods described herein can be adapted to study any brain disease provided that enough training data are available.

To our knowledge, this study was the first to apply a 2-level classification framework based on the 3D HOG feature extraction method to distinguish patients with ASD from healthy controls. We did not rely on 2D HOG as the layer-by-layer slicing method needed can dramatically increase training time and can lead to reduced classification accuracy owing to the separation of the image gradient information from adjacent slices. Of note, in this study we compared 3D to 2D HOG and found that 3D HOG had higher classification accuracy, as demonstrated in Figure 6. Other papers have discussed using the 3D HOG in the medical image domain [70,71]: although the 3D HOG approach may be similar to our approach, we did not concatenate the local HOG features to form a vector representing the entire image. In our framework, we extracted the 3D HOG features for local brain regions and analyzed them individually. In the first-level classification stage, we converted these local features into high-level features with the classification of diseased versus healthy, and then combined these high-level features into a vector. This means the feature dimension input to the final classifier can be considerably reduced, which helps to prevent overfitting. On the contrary, the individual local HOG features can be analyzed further to obtain their respective feature contributions to the ASD classification. These feature contributions actually depict the possibility distribution of the ASD-related brain regions based on the training data. When classifying novel individuals, the feature contributions can be used to discern the most predictive ASD-related brain regions. Importantly, our findings (Tables 3 and 4) also demonstrate that the HOG features outperform SIFT, another widely used local feature, in ASD classification. This is likely due to the ability of the HOG features to cover the entire sMRI image, ensuring that no subtle morphological abnormalities occurring in the brain are overlooked.

In addition to the strengths discussed earlier, our study has several limitations. Specifically, our HOG feature extraction method is based on the artificial division of the brain image with a fixed cell size. The abnormal regions may be located across adjacent cells, and our proposed method considers that such features have the same contribution to the classification

result, which may not entirely reflect the actual grouping complexity. In the future, the HBM framework can be improved by replacing binary classification results like 0 or 1 with fuzzy numbers between 0 and 1 that represent the degree to which the image feature should be classified as a disease-related feature.

Our use of data from 4 ABIDE II sites also presents some challenges. Compared with some other available datasets such as ABIDE I, the ABIDE II datasets and sites are more heterogeneous, which may introduce classification challenges and lead to decreased case versus control classification accuracy. We noted that both Tables 3 and 4 display obvious performance variations between different sites owing to data heterogeneity (eg, differences in scanner types, data collection protocol, demographic information, and disease evaluation). When we applied the HBM method to all the data from the 4 datasets in the 10-fold CV, the resulting classification accuracy reduced to 65% (162/250). This is a common challenge when analyzing multisite data based on neuroimaging techniques. The multisite data heterogeneity makes the classifiers learn site-specific variabilities instead of important information in data themselves. If the data heterogeneous factors are not eliminated, the model performance would not improve even if trained on more data. This is evident in 4 previous studies; the accuracy ranged from 64% to 70% when data from all sites in ABIDE I were integrated [72-75]. In addition, the 2 studies that we compared also used fewer than 4 sites. In our future studies, we will endeavor to reduce the impact of sample site heterogeneity by including scanner parameters and demographic characteristics such as age, sex, and clinical measurements in the analytic models. Another method to address this limitation is through multitask learning, which considers each site as 1 task, and learning of task-shared and task-specific features simultaneously [76,77].

Conclusions

Although ABIDE II study site heterogeneity may have limited case classification accuracy in this study, thus weakening the predictive value of our model, this study nonetheless represents the first steps in developing a classification framework that can distinguish patients with ASD from healthy controls based on the sMRI images that probe the full range of brain regions (subcortical as well as cortical) implicated in ASD. Further development of such sMRI methods—which are more affordable and clinically available than fMRI approaches—to augment the subjective clinical information currently used in the ASD diagnostic process holds much promise, as it could in the future lead to the creation of more accurate and expeditious diagnostic methods.

Acknowledgments

The authors would like to thank Xinyu Guo for providing helpful suggestions and James Ritchie for proofreading the manuscript. This research is partially supported by a grant from the National Science Foundation of China (No. 61772375).

Authors' Contributions

TC, YC, and LL designed the study. TC and YC implemented the algorithm. MY preprocessed the imaging data. MG, TL, HL, and TF gave critical suggestions. TC, YC, TF, MY, and LL drafted the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Calculation process of 3D HOG features.
[DOCX File , 40 KB - medinform_v8i5e15767_app1.docx]

Multimedia Appendix 2
Classification performance comparison for other three datasets.
[DOCX File , 2394 KB - medinform_v8i5e15767_app2.docx]

References

1. Leigh JP, Du J. Brief Report: Forecasting the Economic Burden of Autism in 2015 and 2025 in the United States. J Autism Dev Disord 2015 Dec;45(12):4135-4139. [doi: 10.1007/s10803-015-2521-7] [Medline: 26183723]
2. Simms MD. When autistic behavior suggests a disease other than classic autism. Pediatr Clin North Am 2017 Feb;64(1):127-138. [doi: 10.1016/j.pcl.2016.08.009] [Medline: 27894440]
3. Liptak GS, Benzoni LB, Mruzek DW, Nolan KW, Thingvoll MA, Wade CM, et al. Disparities in diagnosis and access to health services for children with autism: data from the National Survey of Children's Health. J Dev Behav Pediatr 2008 Jun;29(3):152-160. [doi: 10.1097/DBP.0b013e318165c7a0] [Medline: 18349708]
4. Kentrou V, de Veld DM, Mataw KJ, Begeer S. Delayed autism spectrum disorder recognition in children and adolescents previously diagnosed with attention-deficit/hyperactivity disorder. Autism 2019 May;23(4):1065-1072 [FREE Full text] [doi: 10.1177/1362361318785171] [Medline: 30244604]

5. Close HA, Lee L, Kaufmann CN, Zimmerman AW. Co-occurring conditions and change in diagnosis in autism spectrum disorders. Pediatrics 2012 Feb;129(2):e305-e316. [doi: 10.1542/peds.2011-1717] [Medline: 22271695]
6. Arimura H, Magome T, Yamashita Y, Yamamoto D. Computer-aided diagnosis systems for brain diseases in magnetic resonance images. Algorithms 2009;2(3):925-952. [doi: 10.3390/a2030925]
7. El-Dahshan EA, Mohsen HM, Revett K, Salem AM. Computer-aided diagnosis of human brain tumor through MRI: a survey and a new algorithm. Expert Syst Appl 2014;41(11):5526-5545. [doi: 10.1016/j.eswa.2014.01.021]
8. Guo X, Dominick KC, Minai AA, Li H, Erickson CA, Lu LJ. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. Front Neurosci 2017;11:460 [FREE Full text] [doi: 10.3389/fnins.2017.00460] [Medline: 28871217]
9. Price T, Wee CY, Gao W, Shen D. Multiple-Network Classification of Childhood Autism Using Functional Connectivity Dynamics. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. 2014 Presented at: MICCAI'14; September 14-18, 2014; Boston, MA, USA p. 177-184. [doi: 10.1007/978-3-319-10443-0_23]
10. Huang H, Liu X, Jin Y, Lee S, Wee C, Shen D. Enhancing the representation of functional connectivity networks by fusing multi-view information for autism spectrum disorder diagnosis. Hum Brain Mapp 2019 Feb 15;40(3):833-854. [doi: 10.1002/hbm.24415] [Medline: 30357998]
11. Cheng R, Shang Y, Hayes D, Saha SP, Yu G. Noninvasive optical evaluation of spontaneous low frequency oscillations in cerebral hemodynamics. Neuroimage 2012 Sep;62(3):1445-1454. [doi: 10.1016/j.neuroimage.2012.05.069] [Medline: 22659481]
12. Buckner RL, Krienen FM, Yeo BT. Opportunities and limitations of intrinsic functional connectivity MRI. Nat Neurosci 2013 Jul;16(7):832-837. [doi: 10.1038/nn.3423] [Medline: 23799476]
13. Eklund A, Nichols TE, Knutsson H. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc Natl Acad Sci USA 2016 Jul 12;113(28):7900-7905 [FREE Full text] [doi: 10.1073/pnas.1602413113] [Medline: 27357684]
14. Mosconi M, Zwaigenbaum L, Piven J. Structural MRI in autism: Findings and future directions. Clin Neurosci Res 2006;6(3-4):135-144. [doi: 10.1016/j.cnr.2006.06.010]
15. Katuwal GJ, Cahill N, Baum S, Michael AM. The Predictive Power of Structural MRI in Autism Diagnosis. In: Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.: Institute of Electrical and Electronics Engineers; 2015 Presented at: EMBC'15; August 25-29, 2015; Milan, Italy p. 4270-4273. [doi: 10.1109/embc.2015.7319338]
16. Hazlett HC, Gu H, Munsell BC, Kim SH, Styner M, Wolff JJ, IBIS Network, Clinical Sites, Data Coordinating Center, Image Processing Core, Statistical Analysis. Early brain development in infants at high risk for autism spectrum disorder. Nature 2017 Feb 15;542(7641):348-351 [FREE Full text] [doi: 10.1038/nature21369] [Medline: 28202961]
17. Bigler ED, Mortensen S, Neeley ES, Ozonoff S, Krasny L, Johnson M, et al. Superior temporal gyrus, language function, and autism. Dev Neuropsychol 2007;31(2):217-238. [doi: 10.1080/87565640701190841] [Medline: 17488217]
18. Ashburner J, Friston KJ. Voxel-based morphometry--the methods. Neuroimage 2000 Jun;11(6 Pt 1):805-821. [doi: 10.1006/nimg.2000.0582] [Medline: 10860804]
19. Jiao Y, Chen R, Ke X, Chu K, Lu Z, Herskovits EH. Predictive models of autism spectrum disorder based on brain regional cortical thickness. Neuroimage 2010 Apr 1;50(2):589-599. [doi: 10.1016/j.neuroimage.2009.12.047] [Medline: 20026220]
20. Ashburner J, Hutton C, Frackowiak R, Johnsrude I, Price C, Friston K. Identifying global anatomical differences: deformation-based morphometry. Hum Brain Mapp 1998;6(5-6):348-357. [doi: 10.1002/(sici)1097-0193(1998)6:5/6<348::aid-hbm4>3.0.co;2-p] [Medline: 9788071]
21. Bossa M, Zacur E, Olmos S, Alzheimer's Disease Neuroimaging Initiative. Tensor-based morphometry with stationary velocity field diffeomorphic registration: application to ADNI. Neuroimage 2010 Jul 1;51(3):956-969 [FREE Full text] [doi: 10.1016/j.neuroimage.2010.02.061] [Medline: 20211269]
22. Hua X, Leow AD, Parikshak N, Lee S, Chiang M, Toga AW, Alzheimer's Disease Neuroimaging Initiative. Tensor-based morphometry as a neuroimaging biomarker for Alzheimer's disease: an MRI study of 676 AD, MCI, and normal subjects. Neuroimage 2008;43(3):458-469 [FREE Full text] [doi: 10.1016/j.neuroimage.2008.07.013] [Medline: 18691658]
23. Chen R, Jiao Y, Herskovits EH. Structural MRI in autism spectrum disorder. Pediatr Res 2011 May;69(5 Pt 2):63R-68R [FREE Full text] [doi: 10.1203/PDR.0b013e318212c2b3] [Medline: 21289538]
24. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M, Alzheimer's Disease Neuroimaging Initiative. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. Neuroimage 2011 May 15;56(2):766-781. [doi: 10.1016/j.neuroimage.2010.06.013] [Medline: 20542124]
25. Toews M, Wells W, Collins DL, Arbel T. Feature-based morphometry: discovering group-related anatomical patterns. Neuroimage 2010 Feb 1;49(3):2318-2327 [FREE Full text] [doi: 10.1016/j.neuroimage.2009.10.032] [Medline: 19853047]
26. Lee AD, Leow AD, Lu A, Reiss AL, Hall S, Chiang M, et al. 3D pattern of brain abnormalities in Fragile X syndrome visualized using tensor-based morphometry. Neuroimage 2007 Feb 1;34(3):924-938 [FREE Full text] [doi: 10.1016/j.neuroimage.2006.09.043] [Medline: 17161622]
27. Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91-110. [doi: 10.1023/b:visi.0000029664.99615.94]

28. Daliri MR. Automated diagnosis of Alzheimer disease using the scale-invariant feature transforms in magnetic resonance images. J Med Syst 2012 Apr;36(2):995-1000. [doi: 10.1007/s10916-011-9738-6] [Medline: 21584770]
29. Mwangi B, Ebmeier K, Matthews K, Steele J. Multi-centre diagnostic classification of individual structural neuroimaging scans from patients with major depressive disorder. Brain 2012 May;135(Pt 5):1508-1521. [doi: 10.1093/brain/aws084] [Medline: 22544901]
30. Tan L, Chen Y, Maloney TC, Caré MM, Holland SK, Lu LJ. Combined analysis of sMRI and fMRI imaging data provides accurate disease markers for hearing impairment. Neuroimage Clin 2013;3:416-428 [FREE Full text] [doi: 10.1016/j.nicl.2013.09.008] [Medline: 24363991]
31. Chen Y, Storrs J, Tan L, Mazlack LJ, Lee J, Lu LJ. Detecting brain structural changes as biomarker from magnetic resonance images using a local feature based SVM approach. J Neurosci Methods 2014 Jan 15;221:22-31. [doi: 10.1016/j.jneumeth.2013.09.001] [Medline: 24041480]
32. Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2005 Presented at: CVPR'05; June 20-25, 2005; San Diego, CA, USA p. 886-893. [doi: 10.1109/cvpr.2005.177]
33. Zhu Q, Yeh MC, Cheng KT, Avidan S. Fast Human Detection Using a Cascade of Histograms of Oriented Gradients. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2006 Presented at: CVPR'06; June 17-22, 2006; New York, NY, USA p. 1491-1498. [doi: 10.1109/cvpr.2006.119]
34. Li M, Zhang Z, Huang K, Tan T. Estimating the Number of People in Crowded Scenes by MID Based Foreground Segmentation and Head-shoulder Detection. In: Proceedings of the 2008 19th International Conference on Pattern Recognition. 2008 Presented at: ICPR'08; December 8-11, 2008; Tampa, FL, USA. [doi: 10.1109/icpr.2008.4761705]
35. Xie Y, Liu LF, Li CH, Qu YY. Unifying Visual Saliency With HOG Feature Learning for Traffic Sign Detection. In: Proceedings of the 2009 IEEE Intelligent Vehicles Symposium. 2009 Presented at: IVS'09; June 3-5, 2009; Xi'an, China. [doi: 10.1109/ivs.2009.5164247]
36. Overett G, Petersson L. Large Scale Sign Detection Using HOG Feature Variants. In: Proceedings of the 2011 IEEE Intelligent Vehicles Symposium. 2011 Presented at: IVS'11; June 5-9, 2011; Baden-Baden, Germany. [doi: 10.1109/ivs.2011.5940549]
37. Khan S, Cheng H, Matthies D, Sawhney H. 3D Model Based Vehicle Classification in Aerial Imagery. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010 Presented at: CVPR'10; June 13-18, 2010; San Francisco, CA, USA. [doi: 10.1109/cvpr.2010.5539835]
38. Simo-Serra E, Quattoni A, Torras C, Moreno-Noguer F. A Joint Model for 2D and 3D Pose Estimation from a Single Image. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013 Presented at: CVPR'13; June 23-28, 2013; Portland, OR, USA. [doi: 10.1109/cvpr.2013.466]
39. Kobayashi T. BFO Meets HOG: Feature Extraction Based on Histograms of Oriented p.d.f. Gradients for Image Classification. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2013 Presented at: CVPR'13; June 23-28, 2013; Portland, OR, USA. [doi: 10.1109/cvpr.2013.102]
40. Di Martino A, O'Connor D, Chen B, Alaerts K, Anderson JS, Assaf M, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. Sci Data 2017 Mar 14;4:170010 [FREE Full text] [doi: 10.1038/sdata.2017.10] [Medline: 28291247]
41. Meyer D, Leisch F, Hornik K. The support vector machine under test. Neurocomputing 2003;55(1-2):169-186. [doi: 10.1016/s0925-2312(03)00431-4]
42. Focke NK, Helms G, Scheewe S, Pantel PM, Bachmann CG, Dechent P, et al. Individual voxel-based subtype prediction can differentiate progressive supranuclear palsy from idiopathic Parkinson syndrome and healthy controls. Hum Brain Mapp 2011 Nov;32(11):1905-1915. [doi: 10.1002/hbm.21161] [Medline: 21246668]
43. Vemuri P, Gunter JL, Senjem ML, Whitwell JL, Kantarci K, Knopman DS, et al. Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies. Neuroimage 2008 Feb 1;39(3):1186-1197 [FREE Full text] [doi: 10.1016/j.neuroimage.2007.09.073] [Medline: 18054253]
44. Magnin B, Mesrob L, Kinkingnéhun S, Pélégrini-Issac M, Colliot O, Sarazin M, et al. Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI. Neuroradiology 2009 Feb;51(2):73-83. [doi: 10.1007/s00234-008-0463-x] [Medline: 18846369]
45. Ecker C, Rocha-Rego V, Johnston P, Mourao-Miranda J, Marquand A, Daly EM, MRC AIMS Consortium. Investigating the predictive value of whole-brain structural MR scans in autism: a pattern classification approach. Neuroimage 2010 Jan 1;49(1):44-56. [doi: 10.1016/j.neuroimage.2009.08.024] [Medline: 19683584]
46. Calderoni S, Retico A, Biagi L, Tancredi R, Muratori F, Tosetti M. Female children with autism spectrum disorder: an insight from mass-univariate and pattern classification analyses. Neuroimage 2012 Jan 16;59(2):1013-1022. [doi: 10.1016/j.neuroimage.2011.08.070] [Medline: 21896334]
47. Colby JB, Rudie JD, Brown JA, Douglas PK, Cohen MS, Shehzad Z. Insights into multimodal imaging classification of ADHD. Front Syst Neurosci 2012;6:59 [FREE Full text] [doi: 10.3389/fnsys.2012.00059] [Medline: 22912605]

48. Castellani U, Rossato E, Murino V, Bellani M, Rambaldelli G, Perlini C, et al. Classification of schizophrenia using feature-based morphometry. J Neural Transm (Vienna) 2012 Mar;119(3):395-404. [doi: 10.1007/s00702-011-0693-7] [Medline: 21904897]
49. Murphy K. Machine Learning: A Probabilistic Perspective. Cambridge, MA: Mit Press; 2012.
50. Da X, Toledo JB, Zee J, Wolk DA, Xie SX, Ou Y, Alzheimer's Neuroimaging Initiative. Integration and relative value of biomarkers for prediction of MCI to AD progression: spatial patterns of brain atrophy, cognitive scores, APOE genotype and CSF biomarkers. Neuroimage Clin 2014;4:164-173 [FREE Full text] [doi: 10.1016/j.nicl.2013.11.010] [Medline: 24371799]
51. Lao Z, Shen D, Xue Z, Karacali B, Resnick SM, Davatzikos C. Morphological classification of brains via high-dimensional shape transformations and machine learning methods. Neuroimage 2004 Jan;21(1):46-57. [doi: 10.1016/j.neuroimage.2003.09.027] [Medline: 14741641]
52. Etzel JA, Valchev N, Keysers C. The impact of certain methodological choices on multivariate analysis of fMRI data with support vector machines. Neuroimage 2011 Jan 15;54(2):1159-1167. [doi: 10.1016/j.neuroimage.2010.08.050] [Medline: 20817107]
53. Liu M, Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Ensemble sparse classification of Alzheimer's disease. Neuroimage 2012 Apr 2;60(2):1106-1116 [FREE Full text] [doi: 10.1016/j.neuroimage.2012.01.055] [Medline: 22270352]
54. Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D. Random Forest-Based Manifold Learning for Classification of Imaging Data in Dementia. In: Proceedings of the International Workshop on Machine Learning in Medical Imaging. 2011 Presented at: MLMI'11; September 18, 2011; Toronto, Canada p. 159-166. [doi: 10.1007/978-3-642-24319-6_20]
55. Richiardi J, Eryilmaz H, Schwartz S, Vuilleumier P, van de Ville D. Decoding brain states from fMRI connectivity graphs. Neuroimage 2011 May 15;56(2):616-626. [doi: 10.1016/j.neuroimage.2010.05.081] [Medline: 20541019]
56. Acharya UR, Sree SV, Alvin AP, Suri JS. Use of principal component analysis for automatic classification of epileptic EEG activities in wavelet framework. Expert Syst Appl 2012;39(10):9072-9078. [doi: 10.1016/j.eswa.2012.02.040]
57. Lemm S, Blankertz B, Dickhaus T, Müller KR. Introduction to machine learning for brain imaging. Neuroimage 2011 May 15;56(2):387-399. [doi: 10.1016/j.neuroimage.2010.11.004] [Medline: 21172442]
58. Pantelis C, Velakoulis D, McGorry PD, Wood SJ, Suckling J, Phillips LJ, et al. Neuroanatomical abnormalities before and after onset of psychosis: a cross-sectional and longitudinal MRI comparison. Lancet 2003 Jan 25;361(9354):281-288. [doi: 10.1016/S0140-6736(03)12323-9] [Medline: 12559861]
59. Waiter GD, Williams JH, Murray AD, Gilchrist A, Perrett DI, Whiten A. Structural white matter deficits in high-functioning individuals with autistic spectrum disorder: a voxel-based investigation. Neuroimage 2005 Jan 15;24(2):455-461. [doi: 10.1016/j.neuroimage.2004.08.049] [Medline: 15627587]
60. Travers BG, Adluru N, Ennis C, Tromp DP, Destiche D, Doran S, et al. Diffusion tensor imaging in autism spectrum disorder: a review. Autism Res 2012 Oct;5(5):289-313 [FREE Full text] [doi: 10.1002/aur.1243] [Medline: 22786754]
61. Rojas DC, Peterson E, Winterrowd E, Reite ML, Rogers SJ, Tregellas JR. Regional gray matter volumetric changes in autism associated with social and repetitive behavior symptoms. BMC Psychiatry 2006 Dec 13;6:56 [FREE Full text] [doi: 10.1186/1471-244X-6-56] [Medline: 17166273]
62. Pagnozzi AM, Conti E, Calderoni S, Fripp J, Rose SE. A systematic review of structural MRI biomarkers in autism spectrum disorder: A machine learning perspective. Int J Dev Neurosci 2018 Dec;71:68-82. [doi: 10.1016/j.ijdevneu.2018.08.010] [Medline: 30172895]
63. Levman J, Vasung L, MacDonald P, Rowley S, Stewart N, Lim A, et al. Regional volumetric abnormalities in pediatric autism revealed by structural magnetic resonance imaging. Int J Dev Neurosci 2018 Dec;71:34-45. [doi: 10.1016/j.ijdevneu.2018.08.001] [Medline: 30110650]
64. Blakemore SJ. The social brain in adolescence. Nat Rev Neurosci 2008 Apr;9(4):267-277. [doi: 10.1038/nrn2353] [Medline: 18354399]
65. Frith U. Mind blindness and the brain in autism. Neuron 2001 Dec 20;32(6):969-979 [FREE Full text] [doi: 10.1016/s0896-6273(01)00552-9] [Medline: 11754830]
66. Golarai G, Grill-Spector K, Reiss AL. Autism and the development of face processing. Clin Neurosci Res 2006 Oct;6(3):145-160 [FREE Full text] [doi: 10.1016/j.cnr.2006.08.001] [Medline: 18176635]
67. Bastiaansen JA, Thioux M, Nanetti L, van der Gaag C, Ketelaars C, Minderaa R, et al. Age-related increase in inferior frontal gyrus activity and social functioning in autism spectrum disorder. Biol Psychiatry 2011 May 1;69(9):832-838. [doi: 10.1016/j.biopsych.2010.11.007] [Medline: 21310395]
68. Amaral DG, Schumann CM, Nordahl CW. Neuroanatomy of autism. Trends Neurosci 2008 Mar;31(3):137-145. [doi: 10.1016/j.tins.2007.12.005] [Medline: 18258309]
69. Uddin LQ, Menon V. The anterior insula in autism: under-connected and under-examined. Neurosci Biobehav Rev 2009 Sep;33(8):1198-1203 [FREE Full text] [doi: 10.1016/j.neubiorev.2009.06.002] [Medline: 19538989]
70. Serag A, Macnaught G, Denison FC, Reynolds RM, Semple SI, Boardman JP. Histograms of oriented 3D gradients for fully automated fetal brain localization and robust motion correction in 3 T magnetic resonance images. Biomed Res Int 2017;2017:3956363 [FREE Full text] [doi: 10.1155/2017/3956363] [Medline: 28251155]

71. Ghiassian S, Greiner R, Jin P, Brown MR. Using functional or structural magnetic resonance images and personal characteristic data to identify ADHD and autism. PLoS One 2016;11(12):e0166934 [FREE Full text] [doi: 10.1371/journal.pone.0166934] [Medline: 28030565]
72. Nielsen JA, Zielinski BA, Fletcher PT, Alexander AL, Lange N, Bigler ED, et al. Multisite functional connectivity MRI classification of autism: ABIDE results. Front Hum Neurosci 2013;7:599 [FREE Full text] [doi: 10.3389/fnhum.2013.00599] [Medline: 24093016]
73. Abraham A, Milham MP, Di Martino A, Craddock RC, Samaras D, Thirion B, et al. Deriving reproducible biomarkers from multi-site resting-state data: an Autism-based example. Neuroimage 2017 Feb 15;147:736-745. [doi: 10.1016/j.neuroimage.2016.10.045] [Medline: 27865923]
74. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. Neuroimage Clin 2018;17:16-23 [FREE Full text] [doi: 10.1016/j.nicl.2017.08.017] [Medline: 29034163]
75. Dvornek N, Ventola P, Pelphrey K, Duncan J. Identifying Autism from Resting-State fMRI Using Long Short-Term Memory Networks. In: Proceedings of the International Workshop on Machine Learning in Medical Imaging. 2017 Presented at: MLMI'17; September 10, 2017; Quebec City, QC, Canada. [doi: 10.1007/978-3-319-67389-9_42]
76. Wang J, Wang Q, Peng J, Nie D, Zhao F, Kim M, et al. Multi-task diagnosis for autism spectrum disorders using multi-modality features: a multi-center study. Hum Brain Mapp 2017 Jun;38(6):3081-3097 [FREE Full text] [doi: 10.1002/hbm.23575] [Medline: 28345269]
77. Ma Q, Zhang T, Zanetti MV, Shen H, Satterthwaite TD, Wolf DH, et al. Classification of multi-site MR images in the presence of heterogeneity using multi-task learning. Neuroimage Clin 2018;19:476-486. [doi: 10.1016/j.nicl.2018.04.037] [Medline: 29984156]

Abbreviations

AAL: anatomical automatic labeling
ABIDE I: first edition of the Autism Brain Imaging Data Exchange
ABIDE II: second edition of the Autism Brain Imaging Data Exchange
ASD: autism spectrum disorder
AUC: area under the curve
CV: cross-validation
DBM: deformation-based morphometry
DICOM: Digital Imaging and Communications in Medicine
ETH: ETH Zürich
fMRI: functional MRI
HBM: histogram-based morphometry
HOG: histogram of oriented gradients
MNI: Montreal Neurological Institute
MRI: magnetic resonance imaging
OHSU: Oregon Health and Science University
ROI: region of interest
SBM: surface-based morphometry
SIFT: scale-invariant feature transform
sMRI: structural MRI
SU: Stanford University
SVM: support vector machine
TBM: tensor-based morphometry
VBM: voxel-based morphometry

Edited by C Lovis; submitted 06.08.19; peer-reviewed by H Mufti, A Doryab; comments to author 26.10.19; revised version received 01.12.19; accepted 09.02.20; published 08.05.20.

Please cite as:
Chen T, Chen Y, Yuan M, Gerstein M, Li T, Liang H, Froehlich T, Lu L
The Development of a Practical Artificial Intelligence Tool for Diagnosing and Evaluating Autism Spectrum Disorder: Multicenter Study
JMIR Med Inform 2020;8(5):e15767
URL: https://medinform.jmir.org/2020/5/e15767
doi:10.2196/15767
PMID:32041690

Original Paper

An Electronic Clinical Decision Support System for the Management of Low Back Pain in Community Pharmacy: Development and Mixed Methods Feasibility Study

Aron Simon Downie[1,2], MPhil; Mark Hancock[3], PhD; Christina Abdel Shaheed[1], PhD; Andrew J McLachlan[4], PhD; Ahmet Baki Kocaballi[5,6], PhD; Christopher M Williams[7], PhD; Zoe A Michaleff[1,8], PhD; Chris G Maher[1], DMedSc

[1]Institute for Musculoskeletal Health, Sydney School of Public Health, Faculty of Medicine and Health, The University of Sydney, Camperdown, Australia

[2]Faculty of Science and Engineering, Macquarie University, Macquarie Park, Australia

[3]Faculty of Medicine, Health and Human Sciences, Macquarie University, Macquarie Park, Australia

[4]Sydney Pharmacy School, Faculty of Medicine and Health, University of Sydney, Sydney, Australia

[5]Centre for Health Informatics, Australian Institute of Health Innovation, Faculty of Medicine and Health Sciences, Macquarie University, Macquarie Park, Australia

[6]Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia

[7]Hunter New England Population Health, Hunter New England Local Health District, Newcastle, Australia

[8]Institute for Evidence-Based Healthcare, Faculty of Health Sciences and Medicine, Bond University, Gold Coast, Australia

Corresponding Author:
Aron Simon Downie, MPhil
Institute for Musculoskeletal Health
Sydney School of Public Health, Faculty of Medicine and Health
The University of Sydney
Level 10 KGV Building
Missenden Road
Camperdown, 2050
Australia
Phone: 61 2 9850 6382
Email: aron.downie@sydney.edu.au

Abstract

Background: People with low back pain (LBP) in the community often do not receive evidence-based advice and management. Community pharmacists can play an important role in supporting people with LBP as pharmacists are easily accessible to provide first-line care. However, previous research suggests that pharmacists may not consistently deliver advice that is concordant with guideline recommendations and may demonstrate difficulty determining which patients require prompt medical review. A clinical decision support system (CDSS) may enhance first-line care of LBP, but none exists to support the community pharmacist–client consultation.

Objective: This study aimed to develop a CDSS to guide first-line care of LBP in the community pharmacy setting and to evaluate the pharmacist-reported usability and acceptance of the prototype system.

Methods: A cross-platform Web app for the Apple iPad was developed in conjunction with academic and clinical experts using an iterative user-centered design process during interface design, clinical reasoning, program development, and evaluation. The CDSS was evaluated via one-to-one user-testing with 5 community pharmacists (5 case vignettes each). Data were collected via video recording, screen capture, survey instrument (system usability scale), and direct observation.

Results: Pharmacists' agreement with CDSS-generated self-care recommendations was 90% (18/20), with medicines recommendations was 100% (25/25), and with referral advice was 88% (22/25; total 70 recommendations). Pharmacists expressed uncertainty when screening for serious pathology in 40% (10/25) of cases. Pharmacists requested more direction from the CDSS in relation to automated prompts for user input and page navigation. Overall system usability was rated as excellent (mean score 92/100, SD 6.5; 90th percentile compared with similar systems), with acceptance rated as good to excellent.

Conclusions: A novel CDSS (high-fidelity prototype) to enhance pharmacist care of LBP was developed, underpinned by clinical practice guidelines and informed by a multidisciplinary team of experts. User-testing revealed a high level of usability

and acceptance of the prototype system, with suggestions to improve interface prompts and information delivery. The small study sample limits the generalizability of the findings but offers important insights to inform the next stage of system development.

(JMIR Med Inform 2020;8(5):e17203) doi:10.2196/17203

KEYWORDS

low back pain; community pharmacy; decision support systems, clinical

Introduction

Background

Low back pain (LBP) is a major cause of disability worldwide [1], with almost 1 in 5 people reporting LBP at any one time [2]. People with LBP typically consult general practice, allied health, or community pharmacy for advice and management [3,4]. The role of community pharmacists has evolved from dispensing medication and providing medication advice, to include screening and management for a range of health conditions such as minor ailments and chronic health conditions [5-10]. In alignment with this expanding service model, there is interest for community pharmacy to play a greater role in the early management of back pain [11-13]. There are also potential economic benefits for using community pharmacy as an access point for a range of services, with lower patient and health system costs compared with other primary care models [13-15].

Evidence-Practice Gaps in Management of Low Back Pain

Current clinical practice guidelines for the management of LBP recommend first-line care that includes reassurance, advice to stay active and avoid bed rest, and discouraging diagnostic imaging such as plain radiographs unless serious pathology is suspected [3,16]. Despite these guideline recommendations, a substantial gap between evidence and practice still exists [17]. For example, Abdel Shaheed et al [18] reported that community pharmacists and their staff were able to deliver adequate advice on medication use for LBP, but their ability to provide advice on nonpharmacological management such as staying active, avoiding bed rest, and discouraging imaging was inconsistently delivered. The ability to identify presentations that required prompt medical review was also limited for some community pharmacists.

Support for the Community Pharmacist

Clinical decision support systems (CDSSs) are targeted electronic systems that link evidence-based recommendations with the clinical presentation of the individual to improve clinical decision making and support patient engagement with health decisions [19-24]. Recently, CDSSs have been implemented for management of noncancer pain in the primary care setting [25-29], but these do not transfer to the pharmacy setting because of differences in professional training and consultation environment. Pharmacists already have access to CDSSs (eg, management of infection and deprescribing) [30,31], but none exist to support the community pharmacist–client consultation for LBP. Therefore, a CDSS for the early management of LBP in community pharmacy is warranted [32].

The main objective of this study was to develop a CDSS for pharmacists to guide first-line care of LBP in the community pharmacy setting using a mobile data collection system (Apple iPad, Apple Inc). We also sought to evaluate the pharmacist-reported usability and acceptance of the high-fidelity prototype to inform the next stage of CDSS development.

Methods

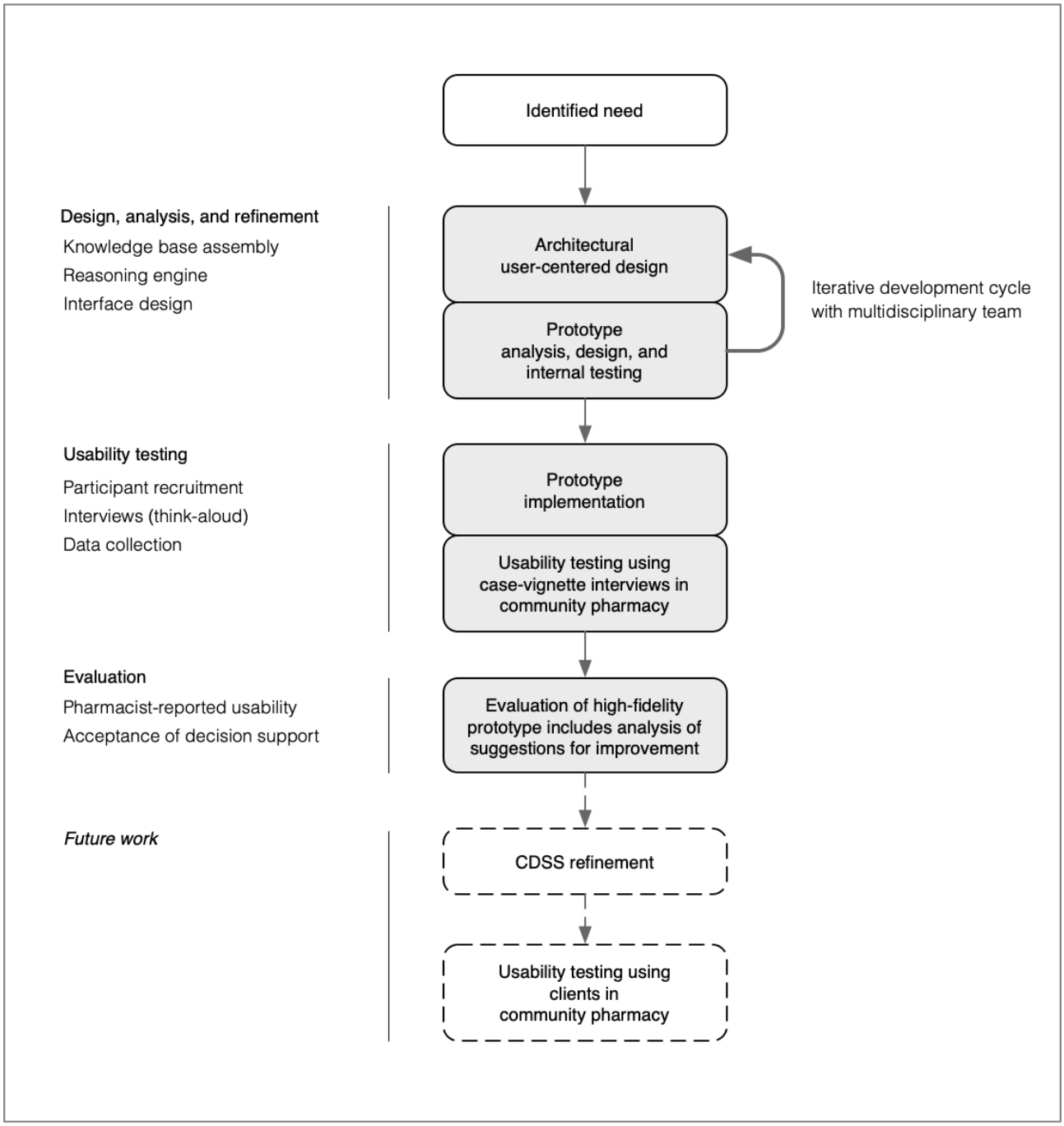
Overview

This study describes the iterative development of a CDSS for the management of pharmacy clients with LBP in community pharmacy. The CDSS was developed by a multidisciplinary team that included two pharmacy academics, a human-computer interaction expert, and four content experts in LBP [33]. Team members were consulted during each stage of development. The CDSS (high-fidelity prototype) was evaluated via a small-scale usability study [34]. The study was approved by University of Sydney Human Research Ethics Committee (2017/027).

User-Centered Design Framework

Development of the CDSS was underpinned by the framework for user-centered design and evaluation of prototypes for clinical information systems (Figure 1) [35]. The framework describes the evolution of a CDSS based around low-cost usability testing methods before future evaluation with real clients in a clinical practice setting. During initial design of the CDSS, input was sought from a range of people involved with community pharmacy, including two pharmacy academic/educators, a community pharmacist, and an industry representative. This approach sought to uncover pharmacist training and procedural constraints that may impact pharmacist decision making for LBP [36], given that a pathway for the contemporary management of LBP specific to the community pharmacy setting does not exist.

Figure 1. Clinical decision support system development based on prototyping and iterative testing (modified from Kushniruk et al). This study is represented by shading. CDSS: clinical decision support system.



Architectural Design, Analysis, and Refinement

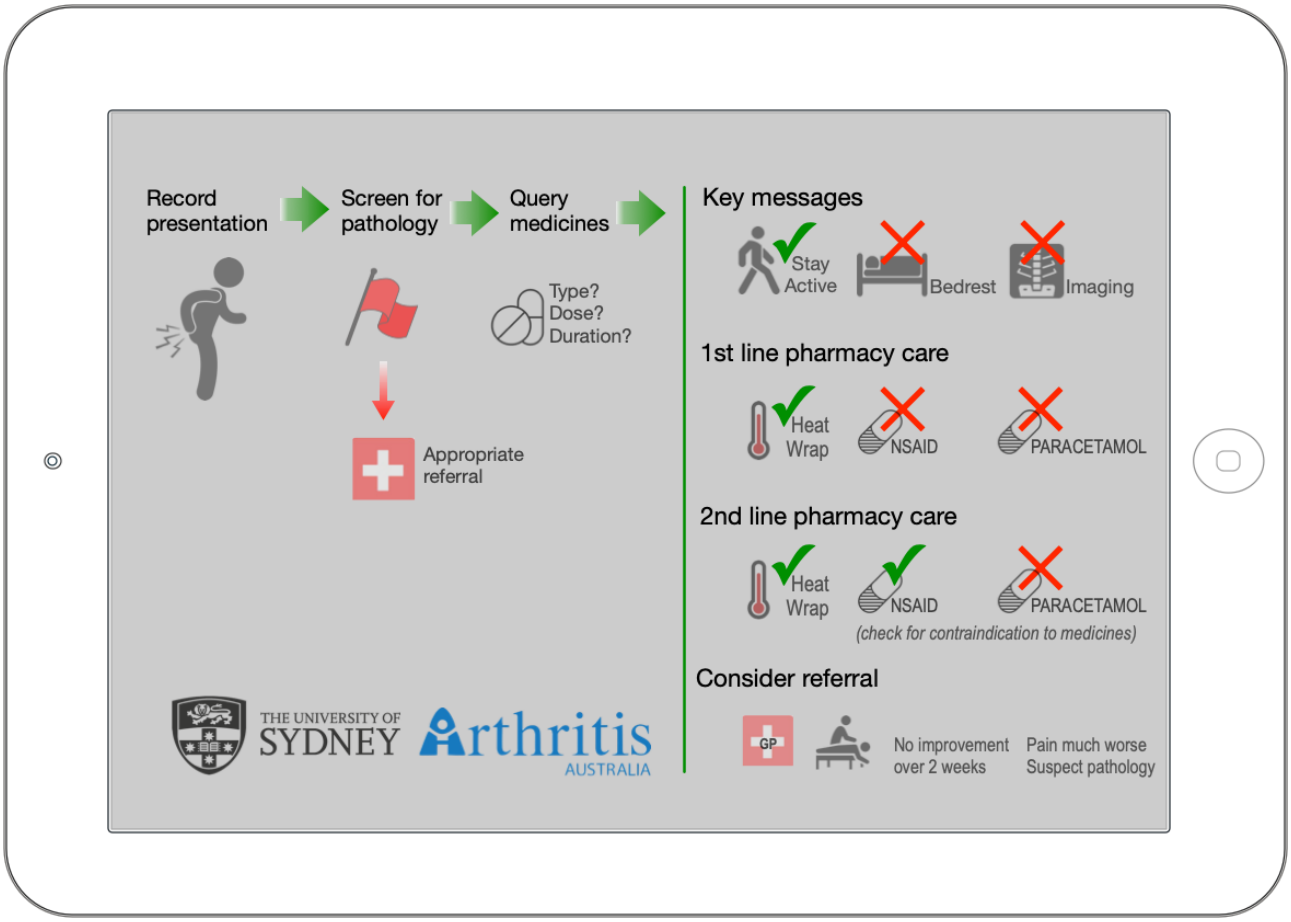
Design goals were informed by Bates et al [37], Khorasani et al [38], and Zikos et al [39] who described features of a decision support system necessary to facilitate integration into clinical practice. The design goals of this CDSS were to (1) support pharmacists to offer simple, clear evidence-based advice to the pharmacy client who presents with LBP; (2) integrate with the pharmacist workflow (eg, consideration of medicines during decision making); (3) maximize time efficiency; and (4) provide a personalized report of recommendations for the pharmacy client.

The CDSS was designed in three components [22]: (1) knowledge base, (2) reasoning engine, and (3) interface (see Multimedia Appendix 1 for further explanation of design process). Briefly, the knowledge base included high level advice for the screening of serious pathology and early management of LBP [3,40-49]. The reasoning engine was coded from the knowledge base then refined using experts in LBP and community pharmacy to consider age, sex, results of screening questions, pain history, and up to three current medicines for LBP (Multimedia Appendix 2). Recommendations for the pharmacist are separated into key messaging for the pharmacy client, suggested medicine use, and referral options. The pharmacist progresses from a landing page (Figure 2), through to history, screening, and advice pages (Multimedia Appendix 3). Data input is via a touch interface (checkboxes, drop-down menus, and free-text input). The pharmacist can accept, modify,

or reject advice generated by the CDSS. Finally, a custom letter for the pharmacy client is generated based on the pharmacist's final recommendation (Multimedia Appendix 4).

The multidisciplinary team of experts were engaged at each stage of design, development, and internal testing. Decision trees were iteratively modified before coding decision logic and programming of the interface (Ionic Framework), then refined through multiple (>10) test cycles. Once the logic and interface were complete, each of the 408 unique decision combinations were checked for accuracy using the Web interface. Similarly, each of the clinical case vignettes (and CDSS-generated client handouts) were tested by the research team for language and accuracy before the interview phase.

Figure 2. Clinical decision support system landing page showing clinical flow and scope of key messages. Tapping anywhere on this screen moves to the "clinical history" page.



Usability Testing

After the completion of the internal testing, the next phase of usability was based on the recommendations of Yen and Bakken [34], where the community pharmacist interacted with the CDSS through a series of 5 case vignettes (system-user-task).

Participants

In all, 5 practising community pharmacists (with 5-27 years of clinical experience), not involved in the initial development phase, from 5 different community pharmacies in the Sydney metropolitan area were each presented with 5 case vignettes during a one-to-one interview. Inclusion criteria required pharmacists to have experience with computer use within pharmacy (eg, computer-based dispensing systems) and be comfortable with tablet computer use (eg, internet browsing). Previous studies suggest that with 5 participants, up to 80% of usability issues can be identified (including up to 100% of major usability issues) when a system is designed for a specific group of users [50,51].

Interview Procedure and Training

Each pharmacist was presented with the same 5 case vignettes role-played by the lead researcher (AD). Clinical scenarios included presentations of both nonserious and serious causes of LBP in adult and elderly populations. Cases 1 and 2 involved nonserious LBP, case 3 involved suspicion of an osteoporotic compression fracture, case 4 presented with nonserious low back and leg pain below the knee, and case 5 presented with LBP and a recent history of cancer (Multimedia Appendix 5).

The interviews were held in a location convenient to the pharmacist, usually in the designated clinical consultation space within the pharmacy. The pharmacist was required to interact with both the "client" (researcher) and the CDSS on an iPad Air (iOS 12.4, Apple Inc). Before beginning the case vignettes, the pharmacist was trained in the operation of the CDSS via interface "walk through." Training also included a brief summary of the evidence underpinning the CDSS, explanation of the pharmacist-client interview process, and how to accept or reject the decision support offered by the CDSS.

The interview was conducted using a think-aloud protocol and employed an active intervention approach [52]. That is, the pharmacist was allowed to ask questions of the researcher during interaction with the "client." Active intervention by the researcher was triggered when the pharmacist was unable to progress through the CDSS, sought clarification when interacting with the interface, or had questions at completion of the case (eg, reflecting upon management decisions generated by the CDSS). All instances of active intervention were logged and evaluated.

Data Collection

Four modes of data collection were used during the interview: (1) think-aloud protocols [53,54] with active intervention approach [52], (2) video/audio recording and screen capture during interaction with the iPad [55,56], (3) direct interview questions at the completion of interaction with the CDSS [26,57], and (4) completion of a survey instrument (system usability scale) [58,59]. The survey instrument was completed by the pharmacist at the end of the interview and without the researcher present (Multimedia Appendix 5).

Evaluation

Evaluation of Usability Testing

Interviews were transcribed verbatim, then independently analyzed by two researchers with assistance by a third (AD, CS and AK) using a directed content analysis methodology, where key concepts from existing usability studies of health information technology methodology were used to inform initial coding categories [34]. Operational definitions for each category were determined based on the specific goals of the CDSS. Any redundant coding categories were collapsed. The analysis of pharmacist sentiment was categorized as "negative," "neutral," or "positive" in consultation with the research team [60]. Frequency of responses were tabulated first by category, then by sentiment (NVivo 12.5, QSR International). Interaction with the iPad was time-stamped to calculate duration spent on each page of the iPad, periods of pharmacist hesitation, and page navigation decisions. Responses to survey instruments were described, and a system usability scale was scored [61].

Level of Acceptance of Clinical Reasoning and Decision Support

At the completion of all case vignettes, each pharmacist was shown an overview of the clinical reasoning engine and then asked to reflect on the logic that informed the recommendation for each case. To quantify the level of acceptance for the core set of recommendations generated by the CDSS (self-care advice, medicines advice, and referral advice), the pharmacist's acceptance (accept/not accept) was logged. Additional advice offered by the pharmacist relating to clinical management was entered in free-text fields on the iPad.

Results

Pharmacist Interview

All pharmacists completed 5 case vignettes on the Apple iPad. Pharmacists were exposed to cases in the same order. The total time taken to role-play all 5 case vignettes (excluding discussion on decision logic or system improvements) ranged from 14 min 35 seconds (Pharmacist #1) to 28 min 4 seconds (Pharmacist #2). Case vignettes that included nonserious LBP required less time (Cases 1 and 2: mean 3 min 40 seconds per case, SD 1 min 8 seconds) than cases that raised suspicion of serious causes of LBP (Cases 3-5: mean 4 min 46 seconds per case, SD 1 min 23 seconds).

Evaluation of Usability Testing

Coding Categories

A total of 162 statements during the 25 interactions between pharmacists and "clients" were logged. Nine coding categories were identified using directed content analysis (*Ease-of-use, Consistency, Visibility, Navigation, Workflow, Content, Understandability, Clarity, and Acceptance)*. For final coding, the categories *Navigation* and *Workflow* were merged, and *Understandability* was defined under *Clarity,* which resulted in seven final categories. Statements were also coded by sentiment (negative, neutral, or positive). Table 1 describes each category, with statement frequency and representative examples. A total of 71 statements related to the CDSS interface, and 91 statements related to clinical information (content) provided by the CDSS.

Table 1. Coding categories with statement frequency and representative examples.

Coding category with subcategory	Sentiment frequency				Representative coded statements with sentiment
	Negative[a]	Neutral[b]	Positive[c]	Total[d]	
Interface					
Ease-of-use: commentary on the simplicity of operation of the CDSS[e]	16	3	7	26	• Negative: (SCREENING page) "It would be a lot easier if it said, 'I've just got some questions I want to ask you, and I just go through them regardless of what you told me." (Pharmacist #5) • Positive: "The App has simple language, it's not complicated, not medical, so that it can be used by everyone. So that's a good thing." (Pharmacist #1)
Consistency: commentary on the consistency of visual language or interaction model	2	7	4	13	• Negative: "I'm pretty sure I did tick 'history of malignancy'. I was surprised that when I ticked that it didn't do what it did do with Betty." (Pharmacist #5) • Positive: (Reads letter) "OK, so it's very similar to the others." (Pharmacist #3)
Visibility: commentary on the visibility of system capabilities and system status and navigational cues within the CDSS	7	6	3	16	• Negative: (ADVICE page) "I didn't notice this one. (points to medicine advice)" (Pharmacist #4) • Positive: "The prompts are there, so it's just something to get used to maneuvering... which isn't very difficult because its laid out quite easily/quite nicely." (Pharmacist #2)
Navigation/workflow: observation and commentary on progression/sequence through the CDSS	2	12	2	16	• Negative: "If we miss one of these pieces of information, does the App ask us to go back?" (Pharmacist #1) • Positive: (HISTORY page) "I really liked this page. I think it's easy to go through." (Pharmacist #3)
Clinical information					
Content: commentary on what information is/is not provided by the CDSS	6	9	12	27	• Negative: (HISTORY page) "Maybe we could add another icon: 'Pregnant or Breastfeeding'" (Pharmacist #4) • Positive: (ADVICE page) "OK, so it actually knows it's sub-therapeutic when I put sub-therapeutic input. That's very good. That's very good." (Pharmacist #2)
Clarity: commentary on the clarity of the information provided by the CDSS	1	9	5	15	• Negative: (SCREENING page) "I know that it's not an infection because they say, 'I fell, and now I've got pain', so it seems like I probably of shouldn't have asked the questions, but I still did because it was still there." (Pharmacist #5) • Positive: (MEDICINES page) "...to recommend Ibuprofen or Aspirin or whatever, then the dose that's required. That's really good. That's really good." (Pharmacist #1)

Coding category with subcategory	Sentiment frequency				Representative coded statements with sentiment
	Negative[a]	Neutral[b]	Positive[c]	Total[d]	
Acceptance: commentary on the clinical value of the CDSS recommendations	6	9	34	49	• Negative: (ADVICE page) "I'm not sure about 'stay active'. I'm not sure it's OK." (Pharmacist #1) • Negative: (ADVICE page) "..and she needs to see someone – like a specialist in this area to find out what is the reason – it is good to have an X-Ray." (Pharmacist #4) • Positive: (Reads letter) "OK. So, stay active. That's really good." (Pharmacist #2) • Positive: (regarding use in practice) "I'd love it... I like the clinical part of my job. I was thinking of having something on pain management plans." (Pharmacist #1)

[a]Negative: negative sentiment.

[b]Neutral: neutral sentiment.

[c]Positive: positive sentiment.

[d]Total: total sentiment count for subcategory.

[e]CDSS: clinical decision support system.

Pharmacists' Statements Related to Interaction With the Interface

The categories *Ease-of-use* and *Visibility* together accounted for 59% (42/71) of statements about the interface with a positive to negative comment ratio of 0.4 (7:16) and 0.4 (3:7), respectively. The majority of statements with negative sentiment involved interaction with the screening page (10/27 statements, Figure 3). The remainder scored with negative sentiment included comments on layout (eg, button position inconsistent) or visibility issues (eg, text size too small).

For example, statements with negative sentiment reported during interaction with the screening page included:

> *I find this part a bit long. I'm always reading through it (risk of spinal inflammation) … maybe it's just me. Maybe I should just read it properly.* [Pharmacist #2]

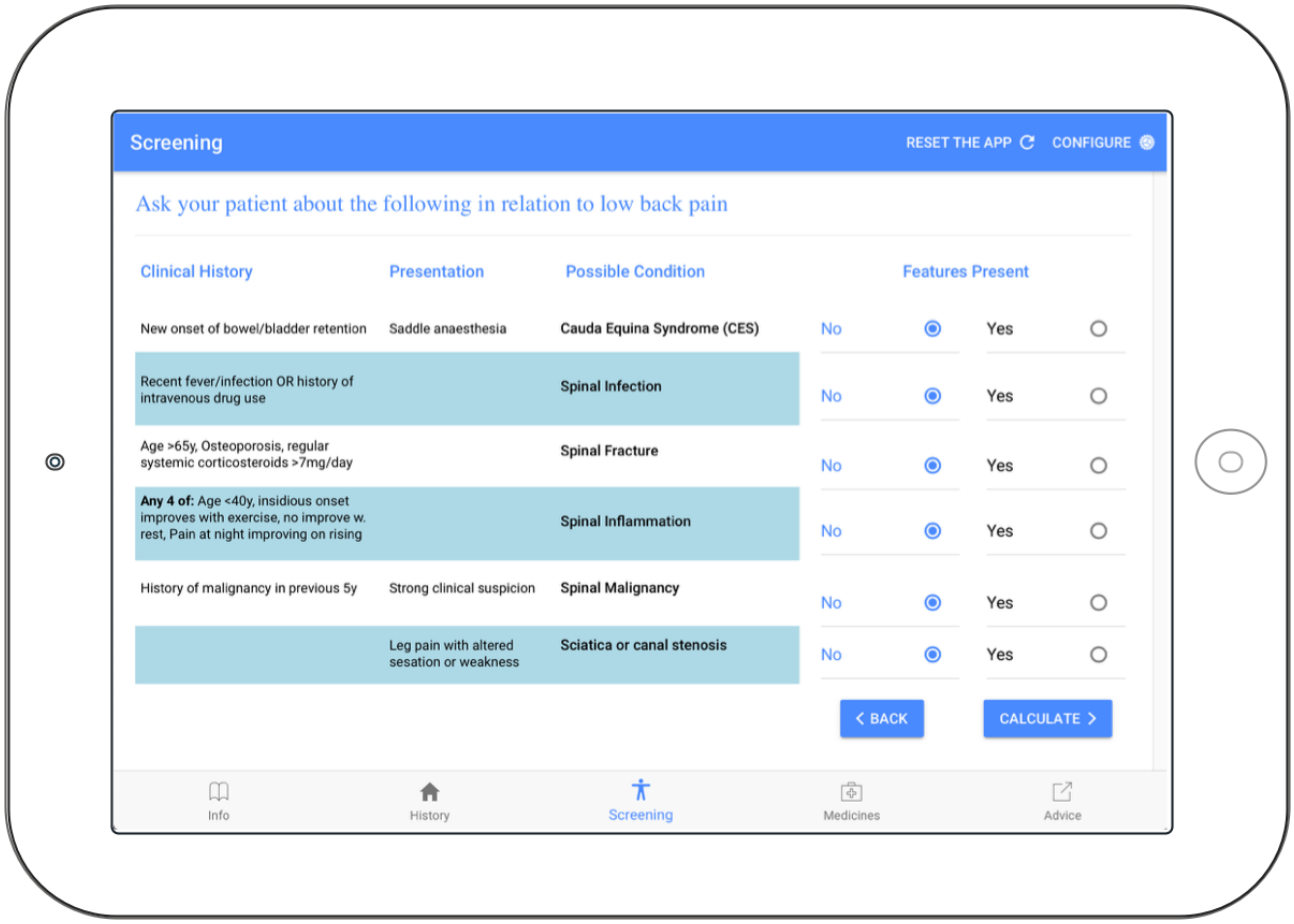
> *[reads from iPad] Leg pain with altered sensation or weakness. So, I guess I didn't see that part... is there a reason that that's here (points to the 2nd column) – Oh, because it's not a clinical history, yep.* [Pharmacist #3]

> *It's just like you are trying to focus on the patient, so you are trying to do two things at once. If the patient was happy for me to pause, 'cause you feel a bit awkward, just processing this whilst the patient is in front of you.* [Pharmacist #4]

> *I think that with this (SCREENING PAGE) is probably the hardest screen here because, like some of the questions I knew, like all of the cases so far, I know that it's not an infection because they say I fell and now I've got pain, so it seems like I probably of shouldn't have asked the questions.* [Pharmacist #5]

Statements with positive sentiment for the interface referenced the simplicity of layout, navigation, and language used (5/16 statements). In addition, statements with positive sentiment were made regarding integration with the pharmacist's workflow (5/16 statements). Pharmacists' statements relating to the operation of the CDSS (eg, "so I just press here," and "then it comes out of the printer?") comprised the majority scored with neutral sentiment (17/28 statements). Queries related to the operation of the CDSS decreased in frequency as each interview progressed.

Figure 3. Clinical decision support system screening page for raising suspicion of a serious cause of low back pain. The "No" response is the default state.



Pharmacists' Statements Related to Clinical Information

The categories *Content* and *Acceptance* together accounted for 84% (76/91) of statements related to clinical information provided by the CDSS, with a positive to negative comment ratio of 2.0 (12:6) and 5.7 (34:6), respectively. The remainder of statements related to *Clarity* of the clinical information provided by the CDSS. Statements with negative sentiment for *Content* (6/27 statements) included request for items absent from history (eg, pharmacists wanted to record current level of pain, whether pregnant or breastfeeding, and history of ulcer). Statements with negative sentiment for *Acceptance* (6/49 statements) included disagreement with, or questioning of, CDSS-generated advice in the categories self-care, medicines, and referral advice. For example:

> *Yes, but we always need to do further investigations to find out… and she needs to see someone – like a specialist in this area to find out what is the reason – it is good to have an X-Ray.* [Pharmacist #4; case 1: nonserious cause of LBP]

> *In my practice, a typical customer that you have just described will usually be on some kind of blood pressure medication – usually – which is why we always tend to recommend paracetamol first.* [Pharmacist #5; case 2: recommendation to begin NSAID therapy]

> *But, from the other point of view, would that narrow the amount of medicine that we recommend? So, from the business-Pharm point of view, would that exclude a lot of products?* [Pharmacist #1; general comment at end of interview]

Statements with positive sentiment for *Content* (12/27 statements) included what pharmacists considered to be the right information displayed at the right time. For example:

> *This is what's really interesting, is what really gives this one the meaning – I like the logic behind it.* [Pharmacist #1; reacting to decision for suspicion of fracture]

> *Yes, we do need to know this.* [Pharmacist #4; points to increased risk of cancer]

> *OK, so it actually knows it's sub-therapeutic when I put sub-therapeutic input. That's very good. That's very good.* [Pharmacist #5]

Statements with positive sentiment for *Acceptance* (34/49 statements) included commentary on the clinical value of information generated by the CDSS in areas of self-care, medicines advice, and referral advice. These statements broadly reflected agreement with advice generated by the CDSS. For example:

> *Most of the time it was very logical, it was very rational and logical… It leads us to the right decision.* [Pharmacist #1]

> *Definitely! Its prompting you to ask questions. I must admit some of those questions we probably don't always ask, but we need to be asking...* [Pharmacist #2]
>
> *OK, so just add on (stay active) Yep. OK. Instead of just seeing the GP straight away. OK. Cool!* [Pharmacist #3; during the selection of self-care advice]
>
> *We excluded some diseases which is good. I found out that she is not taking enough medicine.* [Pharmacist #4]
>
> *Yes, so I would say: using a heat wrap will also help, and I would say take the Voltaren 2-3 times per day with food – yes, it says this already!* [Pharmacist #5]

Similar to pharmacists' questions relating to the interface, comments/questions related to clarification of CDSS-generated advice were scored with neutral sentiment. This type of question also decreased in frequency as each pharmacist moved through the 5 cases.

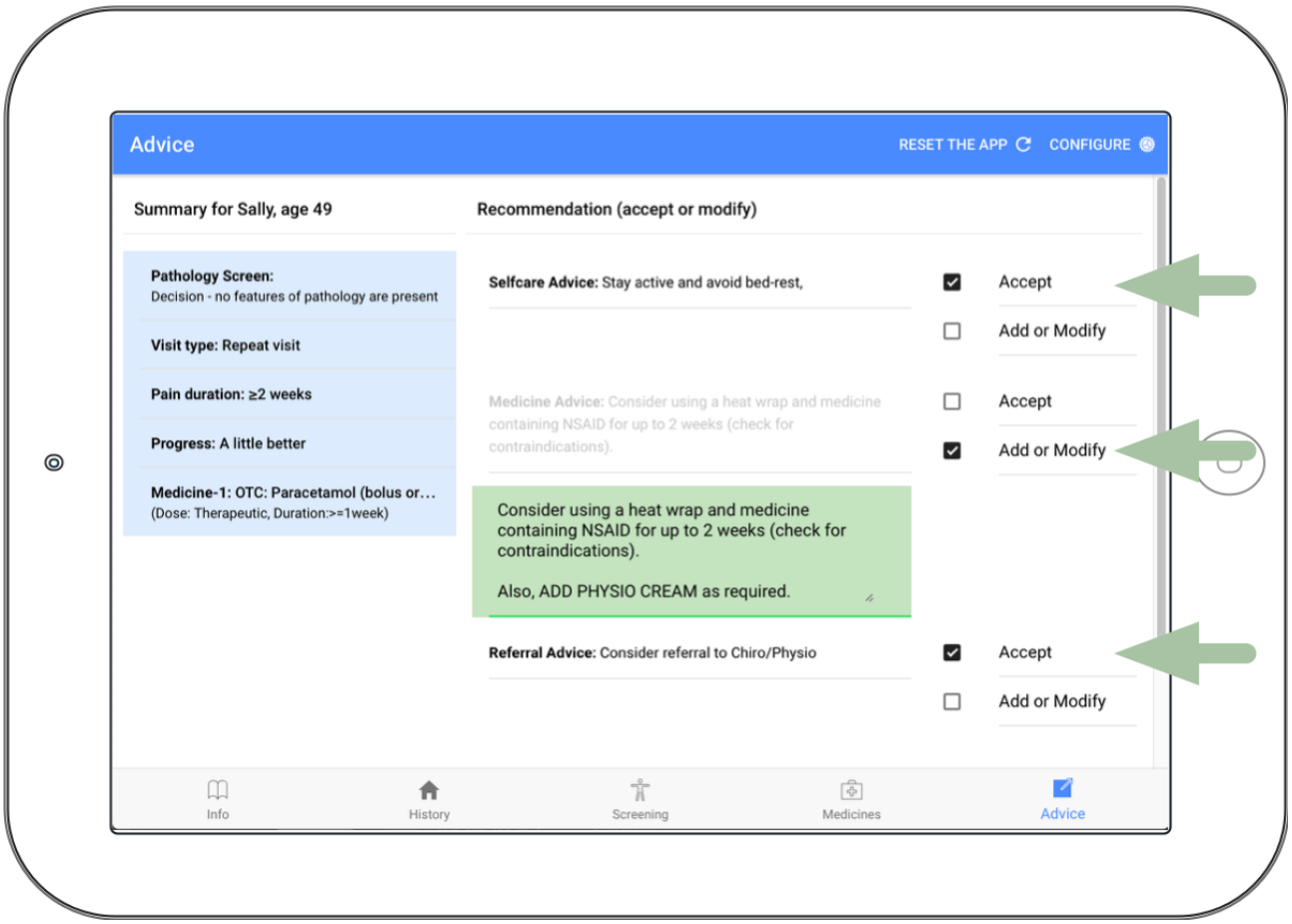
System Usability Scale

The system usability scale [59,61] was administered to each participant at the completion of the interview without the researcher present. Individual usability scores ranged from 82.5 out of 100 to 100 out of 100, which were interpreted as good to excellent usability, respectively [59]. The overall usability score was rated as excellent (mean score 92 out of 100, SD 6.5; 90th percentile compared with similar systems).

Level of Acceptance of Clinical Reasoning and Decision Support

Across the 5 case vignettes, 70 recommendations were generated by the CDSS related to self-care advice, medicine advice, and referral advice. Pharmacists accepted 90% (18/20) of self-care recommendations, 100% (25/25) of medicines recommendations, and 88% (22/25) of referral recommendations. Of those accepted, pharmacists added to the advice for 8% (5/65) of the recommendations generated by the CDSS (eg, Figure 4).

Figure 4. Advice page showing the pharmacist #1 acceptance of self-care advice, the augmentation of medicine advice by the pharmacist ("Also, ADD PHYSIO CREAM as required"), and the acceptance of referral advice (green arrows and green text box).



Discussion

Principal Findings

A CDSS was developed to enhance pharmacist care of LBP, underpinned by clinical practice guidelines and informed by a multidisciplinary team of experts that included consultation with community pharmacy. Community pharmacists rated the overall usability of the high-fidelity protype as good to excellent [59], despite expression of some negative sentiment in relation to guidance in screening for serious causes of LBP and interface inconsistency. There was a high level of acceptance for the advice generated by the CDSS for self-care, medicines, and referral, with pharmacists augmenting advice for a minority (5/65) of recommendations.

Usability

Pharmacists reported a high level of usability based around simple use of language, logical workflow, brief consultation time, ability to customize advice, and convenience of a customized handout for the client. A number of usability issues were raised with regard to interface including page layout, text size, and button placement, which will be considered in the next phase of the CDSS refinement. The screening page (Figure 3) received the majority of negative comments and may reflect nonintuitive interaction with the layout of the screening page and/or lack of familiarity with the screening questions used to raise suspicion of serious causes of LBP. Although education for pharmacists in Australia contains topics on symptom recognition for differential diagnosis [62] and interprofessional referral [63], pharmacists expressed interest for more training on this topic, which is consistent with recommendations of Abdel Shaheed et al [64].

Acceptance of the Clinical Support Provided by the Clinical Decision Support System

All pharmacists agreed that the information provided by the CDSS was applicable to the clinical scenarios presented and could potentially improve client-pharmacist encounters. One pharmacist disagreed with the messaging to avoid imaging and preferred to refer to medical care as a first option for nonserious LBP, but given the small sample, may not be representative of their peers. Pharmacists also commented that the CDSS helped them to ask more questions of the client with LBP and increased management options for LBP beyond their usual advice. However, it is unclear if the advice delivered by the CDSS in this setting would be superior to usual pharmacy care for LBP.

Guidance for Pain Management in Community Pharmacy

Pharmacists commented that they appreciated guidance provided by the CDSS in relation to management, particularly for options beyond medicines advice. This aligns with recommendations of Abdel Shaheed et al [65] and others [11] on the potential benefit of tools/guidelines to support pharmacists when managing clients with LBP. Pharmacists also reflected on the current general lack of guidance to manage pain within pharmacy compared with the promotion and availability of management tools for other health conditions [5,7,8,14,66]. This view is consistent with results from a recent study by Abdel Shaheed et al [65] who found that pharmacists were receptive to implementing a disease state management program for LBP. One area highlighted by pharmacists was the lack of operational knowledge in relation to screening clients for serious causes of LBP, which has been highlighted previously [18]. Abdel Shaheed et al also found that pharmacists had both the willingness and capacity to increase knowledge in this area [32,64]. One goal of a training module integrated into the next version of the CDSS would be to empower the pharmacist with the skillset to raise suspicion of potentially serious underlying pathology, then inform clients of options for prompt medical review [43,45].

Limitations

The small sample size may not be adequate to capture the full range of pharmacists' views or usability issues thus limiting the generalizability of the results [67], particularly with regard to the level of acceptance. However, the sample size was appropriate for this stage of CDSS development [34,51]. That is, it was sufficient to identify major usability issues (eg, when screening for risk of serious disease), that the CDSS interface could be navigated with minimal training, and that decisions generated were logical and easy for the pharmacist to apply (in a simulated scenario). The method used to assess usability (think aloud with active intervention) may have enhanced task performance through researcher-induced bias [68] but allowed greater insight into the sections of the CDSS that required further development [52]. Another source of bias that may have enhanced task performance was the nonrandomized order of cases (case complexity was greater later in interview). This stage of CDSS development was to finalize design elements in the community pharmacy setting before testing with real clients [34]. In its current design, the CDSS does not integrate with existing electronic record systems in pharmacy, which would be necessary before advanced testing and would increase the chance of adoption by pharmacists [69]. One approach would be integration with existing disease state management systems [70], which was also suggested by pharmacists during testing.

Comparison With Prior Work

This CDSS is the first tool that the authors are aware of to assist community pharmacists in first-line care for people with LBP. Other electronic decision support systems have been targeted at the primary care setting for the management of LBP [25] and chronic pain [26,27]. This CDSS differs from existing systems in that it aims to empower the pharmacist to offer evidence-based first-line care beyond medicines advice, and stepped referral options to allied health, primary, or emergency care based on presentation or symptom progression. The opportunity to enhance the pharmacist-client interaction, identified as lacking in other systems [27,71], has been built into this CDSS by allowing the pharmacist to modify management advice then provide a customized handout for the client.

Future Direction

The next phase is to modify the CDSS with lessons learned from this usability study, then reevaluate during the next level of system development (integration with clients into the pharmacy setting) [34]. The CDSS will also be evaluated with respect to the credibility of advice and satisfaction with care from the perspective of clients with LBP. An education module on the evidence-based management of LBP could be delivered to pharmacists in conjunction with training for the CDSS, which would assist with knowledge of screening for pathology and give context to the guideline-based care options suggested by the CDSS. Future studies may establish if pharmacist training during the use of a CDSS within the clinical encounter improves both pharmacist and pharmacy client satisfaction with care.

Conclusions

Despite many years of clinical guidelines for the management of LBP, significant evidence-to-practice gaps remain. This CDSS has been designed to provide a unique opportunity for community pharmacists to provide simple evidence-based advice for clients who present with LBP. Importantly the CDSS offers key messages of reassurance, to remain active, to use medicines appropriately, and to avoid inappropriate imaging.

Acknowledgments

This study was supported by the 2017 Arthritis Australia Grant in Aid (Zimmer Biomet Australia Grant). The authors would like to acknowledge the support of the pharmacists who contributed during development of the CDSS or were participants in this study. The authors also thank professors Bart Koes, Ric Day, and Ian Harris who provided input during the development of the CDSS. A National Health and Medical Research Council Program Grant (ID APP1113532) and Research Fellowship (ID APP1103022) supports CM.

Authors' Contributions

AD, CM, MH, CS, AM, CW and ZM designed the study. AD and AK were responsible for interface design and coding with revision from all authors. AD was responsible for the acquisition of data. AD, CS, and AK analyzed interview transcripts. AD wrote the first draft of the manuscript. All authors critically reviewed the manuscript and reviewed the final draft before submission.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Clinical decision support system design methodology.
[DOCX File , 23 KB - medinform_v8i5e17203_app1.docx]

Multimedia Appendix 2
Clinical decision support system decision tree.
[PDF File (Adobe PDF File), 402 KB - medinform_v8i5e17203_app2.pdf]

Multimedia Appendix 3
Clinical decision support system data entry screens.
[PDF File (Adobe PDF File), 2159 KB - medinform_v8i5e17203_app3.pdf]

Multimedia Appendix 4
Client handout example generated by clinical decision support system.
[PDF File (Adobe PDF File), 98 KB - medinform_v8i5e17203_app4.pdf]

Multimedia Appendix 5
Interview guide.
[PDF File (Adobe PDF File), 396 KB - medinform_v8i5e17203_app5.pdf]

References

1. Smith E, Hoy DG, Cross M, Vos T, Naghavi M, Buchbinder R, et al. The global burden of other musculoskeletal disorders: estimates from the Global Burden of Disease 2010 study. Ann Rheum Dis 2014 Aug;73(8):1462-1469. [doi: 10.1136/annrheumdis-2013-204680] [Medline: 24590181]
2. Hoy D, Bain C, Williams G, March L, Brooks P, Blyth F, et al. A systematic review of the global prevalence of low back pain. Arthritis Rheum 2012 Jun;64(6):2028-2037 [FREE Full text] [doi: 10.1002/art.34347] [Medline: 22231424]
3. Qaseem A, Wilt T, McLean R, Forciea M, Clinical Guidelines Committee of the American College of Physicians. Noninvasive treatments for acute, subacute, and chronic low back pain: A clinical practice guideline from the American College of Physicians. Ann Intern Med 2017 Apr 4;166(7):514-530. [doi: 10.7326/M16-2367] [Medline: 28192789]
4. Walker BF, Muller R, Grant WD. Low back pain in Australian adults. health provider utilization and care seeking. J Manipulative Physiol Ther 2004 Jun;27(5):327-335. [doi: 10.1016/j.jmpt.2004.04.006] [Medline: 15195040]
5. Kaczorowski J, Chambers LW, Dolovich L, Paterson JM, Karwalajtys T, Gierman T, et al. Improving cardiovascular health at population level: 39 community cluster randomised trial of Cardiovascular Health Awareness Program (CHAP). Br Med J 2011 Feb 7;342:d442 [FREE Full text] [doi: 10.1136/bmj.d442] [Medline: 21300712]

6. Giaccone M, Baratta F, Allais G, Brusa P. Prevention, education and information: the role of the community pharmacist in the management of headaches. Neurol Sci 2014 May;35(Suppl 1):1-4. [doi: 10.1007/s10072-014-1732-6] [Medline: 24867826]
7. Ali M, Schifano F, Robinson P, Phillips G, Doherty L, Melnick P, et al. Impact of community pharmacy diabetes monitoring and education programme on diabetes management: a randomized controlled study. Diabet Med 2012 Sep;29(9):e326-e333. [doi: 10.1111/j.1464-5491.2012.03725.x] [Medline: 22672148]
8. Brown D, Portlock J, Rutter P. Review of services provided by pharmacies that promote healthy living. Int J Clin Pharm 2012 Jun;34(3):399-409. [doi: 10.1007/s11096-012-9634-2] [Medline: 22527479]
9. Greer N, Bolduc J, Geurkink E, Rector T, Olson K, Koeller E, et al. Pharmacist-led chronic disease management: A systematic review of effectiveness and harms compared with usual care. Ann Intern Med 2016 Jul 5;165(1):30-40. [doi: 10.7326/M15-3058] [Medline: 27111098]
10. Légat L, van Laere S, Nyssen M, Steurbaut S, Dupont AG, Cornu P. Clinical decision support systems for drug allergy checking: Systematic review. J Med Internet Res 2018 Sep 7;20(9):e258 [FREE Full text] [doi: 10.2196/jmir.8206] [Medline: 30194058]
11. Silcock J, Moffett JK, Edmondson H, Waddell G, Burton AK. Do community pharmacists have the attitudes and knowledge to support evidence based self-management of low back pain? BMC Musculoskelet Disord 2007 Jan 31;8:10 [FREE Full text] [doi: 10.1186/1471-2474-8-10] [Medline: 17266748]
12. Mishriky J, Stupans I, Chan V. Expanding the role of Australian pharmacists in community pharmacies in chronic pain management - a narrative review. Pharm Pract (Granada) 2019;17(1):1410 [FREE Full text] [doi: 10.18549/PharmPract.2019.1.1410] [Medline: 31015881]
13. Bennett MI, Bagnall A, Raine G, Closs SJ, Blenkinsopp A, Dickman A, et al. Educational interventions by pharmacists to patients with chronic pain: systematic review and meta-analysis. Clin J Pain 2011 Sep;27(7):623-630. [doi: 10.1097/AJP.0b013e31821b6be4] [Medline: 21610491]
14. Crealey GE, McElnay JC, Maguire TA, O'Neill C. Costs and effects associated with a community pharmacy-based smoking-cessation programme. Pharmacoeconomics 1998 Sep;14(3):323-333. [doi: 10.2165/00019053-199814030-00008] [Medline: 10186470]
15. San-Juan-Rodriguez A, Newman T, Hernandez I, Swart E, Klein-Fedyshin M, Shrank W, et al. Impact of community pharmacist-provided preventive services on clinical, utilization, and economic outcomes: An umbrella review. Prev Med 2018 Oct;115:145-155. [doi: 10.1016/j.ypmed.2018.08.029] [Medline: 30145351]
16. National Institute for Health and Care Excellence (NICE). Low Back Pain and Sciatica in Over 16s: Assessment and Management URL: https://pathways.nice.org.uk/pathways/low-back-pain-and-sciatica#content=view-node%3Anodes-imaging [accessed 2019-11-15]
17. Foster NE, Anema JR, Cherkin D, Chou R, Cohen SP, Gross DP, Lancet Low Back Pain Series Working Group. Prevention and treatment of low back pain: evidence, challenges, and promising directions. Lancet 2018 Jun 9;391(10137):2368-2383. [doi: 10.1016/S0140-6736(18)30489-6] [Medline: 29573872]
18. Abdel Shaheed C, McFarlane B, Maher C, Williams K, Bergin J, Matthews A, et al. Investigating the primary care management of low back pain: a simulated patient study. J Pain 2016 Jan;17(1):27-35. [doi: 10.1016/j.jpain.2015.09.010] [Medline: 26456675]
19. van der Weijden T, Boivin A, Burgers J, Schünemann HJ, Elwyn G. Clinical practice guidelines and patient decision aids. An inevitable relationship. J Clin Epidemiol 2012 Jun;65(6):584-589. [doi: 10.1016/j.jclinepi.2011.10.007] [Medline: 22297117]
20. Stacey D, Légaré F, Lewis K, Barry M, Bennett C, Eden K, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev 2017 Apr 12;4:CD001431 [FREE Full text] [doi: 10.1002/14651858.CD001431.pub5] [Medline: 28402085]
21. Bright T, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med 2012 Jul 3;157(1):29-43. [doi: 10.7326/0003-4819-157-1-201207030-00450] [Medline: 22751758]
22. Berner E, La Lande LT. Overview of clinical decision support systems. In: Berner E, editor. Clinical Decision Support Systems. Switzerland: Springer; 2016:1-17.
23. Bryan C, Boren S. The use and effectiveness of electronic clinical decision support tools in the ambulatory/primary care setting: a systematic review of the literature. Inform Prim Care 2008;16(2):79-91 [FREE Full text] [doi: 10.14236/jhi.v16i2.679] [Medline: 18713524]
24. Mickan S, Tilson J, Atherton H, Roberts N, Heneghan C. Evidence of effectiveness of health care professionals using handheld computers: a scoping review of systematic reviews. J Med Internet Res 2013 Oct 28;15(10):e212 [FREE Full text] [doi: 10.2196/jmir.2530] [Medline: 24165786]
25. Peiris D, Williams C, Holbrook R, Lindner R, Reeve J, Das A, et al. A web-based clinical decision support tool for primary health care management of back pain: development and mixed methods evaluation. JMIR Res Protoc 2014 Apr 2;3(2):e17 [FREE Full text] [doi: 10.2196/resprot.3071] [Medline: 24694921]

26. Trafton J, Martins S, Michel M, Lewis E, Wang D, Combs A, et al. Evaluation of the acceptability and usability of a decision support system to encourage safe and effective use of opioid therapy for chronic, noncancer pain by primary care providers. Pain Med 2010 Apr;11(4):575-585. [doi: 10.1111/j.1526-4637.2010.00818.x] [Medline: 20202142]
27. Smith MY, DePue JD, Rini C. Computerized decision-support systems for chronic pain management in Primary Care. Pain Med 2007;8(suppl 3):S155-S166. [doi: 10.1111/j.1526-4637.2007.00278.x]
28. Guenter D, Abouzahra M, Schabort I, Radhakrishnan A, Nair K, Orr S, et al. Design process and utilization of a novel clinical decision support system for neuropathic pain in primary care: Mixed methods observational study. JMIR Med Inform 2019 Sep 30;7(3):e14141 [FREE Full text] [doi: 10.2196/14141] [Medline: 31573946]
29. Nicol A, Hurley R, Benzon H. Alternatives to opioids in the pharmacologic management of chronic pain syndromes: A narrative review of randomized, controlled, and blinded clinical trials. Anesth Analg 2017 Nov;125(5):1682-1703 [FREE Full text] [doi: 10.1213/ANE.0000000000002426] [Medline: 29049114]
30. Curtain C, Peterson GM. Review of computerized clinical decision support in community pharmacy. J Clin Pharm Ther 2014 Aug;39(4):343-348. [doi: 10.1111/jcpt.12168] [Medline: 24806361]
31. Monteiro L, Maricoto T, Solha I, Ribeiro-Vaz I, Martins C, Monteiro-Soares M. Reducing potentially inappropriate prescriptions for older patients using computerized decision support tools: Systematic review. J Med Internet Res 2019 Nov 14;21(11):e15385 [FREE Full text] [doi: 10.2196/15385] [Medline: 31724956]
32. Abdel Shaheed C, Maher CG, Williams KA, McLachlan AJ. Pharmacists' views on implementing a disease state management program for low back pain. Aust J Prim Health 2016;22(3):211-217. [doi: 10.1071/PY14116] [Medline: 25719762]
33. Mugisha A, Babic A, Wakholi P, Tylleskär T. High-fidelity prototyping for mobile electronic data collection forms through design and user evaluation. JMIR Hum Factors 2019 Mar 22;6(1):e11852 [FREE Full text] [doi: 10.2196/11852] [Medline: 30900995]
34. Yen P, Bakken S. Review of health information technology usability study methodologies. J Am Med Inform Assoc 2012;19(3):413-422 [FREE Full text] [doi: 10.1136/amiajnl-2010-000020] [Medline: 21828224]
35. Kushniruk A, Monkman H, Borycki E, Kannry J. User-centered designevaluation of clinical information systems: a usability engineering perspective. In: Patel VL, Kannampallil TG, Kaufman DR, editors. Cognitive Informatics for Biomedicine: Human Computer Interaction in Healthcare. Switzerland: Springer International Publishing; 2015:141-162.
36. Karsh B. Agency for Healthcare Research and Quality. 2009. Clinical Practice Improvement and Redesign: How Change in Workflow Can Be Supported by Clinical Decision Support URL: https://digital.ahrq.gov/sites/default/files/docs/biblio/09-0054-EF-Updated_0.pdf [accessed 2019-11-15]
37. Bates D, Kuperman G, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc 2003;10(6):523-530 [FREE Full text] [doi: 10.1197/jamia.M1370] [Medline: 12925543]
38. Khorasani R, Hentel K, Darer J, Langlotz C, Ip IK, Manaker S, et al. Ten commandments for effective clinical decision support for imaging: enabling evidence-based practice to improve quality and reduce waste. AJR Am J Roentgenol 2014 Nov;203(5):945-951. [doi: 10.2214/AJR.14.13134] [Medline: 25341131]
39. Zikos D, DeLellis N. CDSS-RM: a clinical decision support system reference model. BMC Med Res Methodol 2018 Nov 16;18(1):137 [FREE Full text] [doi: 10.1186/s12874-018-0587-6] [Medline: 30445910]
40. Abdel Shaheed C, Maher CG, Williams KA, McLachlan AJ. Interventions available over the counter and advice for acute low back pain: systematic review and meta-analysis. J Pain 2014 Jan;15(1):2-15. [doi: 10.1016/j.jpain.2013.09.016] [Medline: 24373568]
41. Chou R, Deyo R, Friedly J, Skelly A, Hashimoto R, Weimer M, et al. Nonpharmacologic therapies for low back pain: A systematic review for an American College of Physicians clinical practice guideline. Ann Intern Med 2017 Apr 4;166(7):493-505. [doi: 10.7326/M16-2459] [Medline: 28192793]
42. Chou R, Deyo R, Friedly J, Skelly A, Weimer M, Fu R, et al. Systemic pharmacologic therapies for low back pain: A systematic review for an American College of Physicians clinical practice guideline. Ann Intern Med 2017 Apr 4;166(7):480-492. [doi: 10.7326/M16-2458] [Medline: 28192790]
43. Downie A, Williams C, Henschke N, Hancock M, Ostelo R, de Vet HC, et al. Red flags to screen for malignancy and fracture in patients with low back pain: systematic review. Br Med J 2013 Dec 11;347:f7095 [FREE Full text] [doi: 10.1136/bmj.f7095] [Medline: 24335669]
44. Grossman JM, Gordon R, Ranganath VK, Deal C, Caplan L, Chen W, et al. American College of Rheumatology 2010 recommendations for the prevention and treatment of glucocorticoid-induced osteoporosis. Arthritis Care Res (Hoboken) 2010 Nov;62(11):1515-1526 [FREE Full text] [doi: 10.1002/acr.20295] [Medline: 20662044]
45. Maher C, Underwood M, Buchbinder R. Non-specific low back pain. Lancet 2017 Feb 18;389(10070):736-747. [doi: 10.1016/S0140-6736(16)30970-9] [Medline: 27745712]
46. Mathieson H, Marzo-Ortega H. Axial spondyloarthritis: diagnosis and management. Prescriber 2015;25(23-24):32-36. [doi: 10.1002/psb.1290]
47. Mathieson S, Kasch R, Maher CG, Pinto RZ, McLachlan AJ, Koes BW, et al. Combination drug therapy for the management of low back pain and sciatica: Systematic review and meta-analysis. J Pain 2019 Jan;20(1):1-15. [doi: 10.1016/j.jpain.2018.06.005] [Medline: 30585164]

48. McRae M, Hancock M. Adults attending private physiotherapy practices seek diagnosis, pain relief, improved function, education and prevention: a survey. J Physiother 2017 Oct;63(4):250-256 [FREE Full text] [doi: 10.1016/j.jphys.2017.08.002] [Medline: 28967562]
49. Verbeek J, Sengers M, Riemens L, Haafkens J. Patient expectations of treatment for back pain: a systematic review of qualitative and quantitative studies. Spine (Phila Pa 1976) 2004 Oct 15;29(20):2309-2318. [doi: 10.1097/01.brs.0000142007.38256.7f] [Medline: 15480147]
50. Johnson CM, Johnson TR, Zhang J. A user-centered framework for redesigning health care interfaces. J Biomed Inform 2005 Feb;38(1):75-87 [FREE Full text] [doi: 10.1016/j.jbi.2004.11.005] [Medline: 15694887]
51. Nielsen J. Estimating the number of subjects needed for a thinking aloud test. Int J Hum Comput Stud 1994;41(3):385-397. [doi: 10.1006/ijhc.1994.1065]
52. Alhadreti O, Mayhew P. To intervene or not to intervene: an investigation of three think-aloud protocols in usability testing. J Usability Stud 2017;12(3):111-132 [FREE Full text]
53. Beuscart-Zéphir MC, Brender J, Beuscart R, Ménager-Depriester I. Cognitive evaluation: how to assess the usability of information technology in healthcare. Comput Methods Programs Biomed 1997 Sep;54(1-2):19-28. [doi: 10.1016/s0169-2607(97)00030-8] [Medline: 9290916]
54. Jaspers M, Steen T, van den Bos C, Geenen M. The think aloud method: a guide to user interface design. Int J Med Inform 2004 Nov;73(11-12):781-795. [doi: 10.1016/j.ijmedinf.2004.08.003] [Medline: 15491929]
55. Li AC, Kannry JL, Kushniruk A, Chrimes D, McGinn TG, Edonyabo D, et al. Integrating usability testing and think-aloud protocol analysis with 'near-live' clinical simulations in evaluating clinical decision support. Int J Med Inform 2012 Nov;81(11):761-772. [doi: 10.1016/j.ijmedinf.2012.02.009] [Medline: 22456088]
56. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform 2004 Feb;37(1):56-76 [FREE Full text] [doi: 10.1016/j.jbi.2004.01.003] [Medline: 15016386]
57. Bastien J. Usability testing: a review of some methodological and technical aspects of the method. Int J Med Inform 2010 Apr;79(4):e18-e23. [doi: 10.1016/j.ijmedinf.2008.12.004] [Medline: 19345139]
58. McLellan S, Muddimer A, Peres C. The effect of experience on system usability scale ratings. J Usability Stud 2012;7(2):56-67 [FREE Full text]
59. Bangor A, Kortum PT, Miller JT. An empirical evaluation of the System Usability Scale. Int J Hum Comput Interact 2008;24(6):574-594. [doi: 10.1080/10447310802205776]
60. Hsieh H, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res 2005 Nov;15(9):1277-1288. [doi: 10.1177/1049732305276687] [Medline: 16204405]
61. Lewis JR. The System Usability Scale: past, present, and future. Int J Hum Comput Interact 2018;34(7):577-590. [doi: 10.1080/10447318.2018.1455307]
62. Marriott JL, Nation RL, Roller L, Costelloe M, Galbraith K, Stewart P, et al. Pharmacy education in the context of Australian practice. Am J Pharm Educ 2008 Dec 15;72(6):131 [FREE Full text] [doi: 10.5688/aj7206131] [Medline: 19325951]
63. University of Sydney. 2019. Unit of study: PHAR3826: Musculoskeletal, Dermatological and Senses URL: https://sydney.edu.au/courses/units-of-study/2019/phar/phar3826.html [accessed 2019-11-15]
64. Abdel Shaheed C, Maher CG, Mak W, Williams KA, McLachlan AJ. Knowledge and satisfaction of pharmacists attending an educational workshop on evidence-based management of low back pain. Aust J Prim Health 2015;21(2):126-131. [doi: 10.1071/PY14020] [Medline: 24802263]
65. Abdel Shaheed C, Maher CG, Mak W, Williams KA, McLachlan AJ. The effects of educational interventions on pharmacists' knowledge, attitudes and beliefs towards low back pain. Int J Clin Pharm 2015 Aug;37(4):616-625. [doi: 10.1007/s11096-015-0112-5] [Medline: 25851502]
66. Hanna A, White L, Yanamandram V. Patients' willingness to pay for diabetes disease state management services in Australian community pharmacies. Intl J of Pharm Health Mrkt 2010;4(4):339-354. [doi: 10.1108/17506121011095191]
67. Faulkner L. Beyond the five-user assumption: benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput 2003 Aug;35(3):379-383. [doi: 10.3758/bf03195514] [Medline: 14587545]
68. Zhao T, McDonald S, Edwards H. The impact of two different think-aloud instructions in a usability test: a case of just following orders? Behav Inform Technol 2014;33(2):163-183. [doi: 10.1080/0144929X.2012.708786]
69. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: Critical analysis. JMIR Med Inform 2018 Apr 18;6(2):e24 [FREE Full text] [doi: 10.2196/medinform.8912] [Medline: 29669706]
70. GuildCare: Pharmacy software for professional services. Sydney, Australia: GuildLink; 2019. URL: http://www.guildlink.com.au/guildcare/ [accessed 2019-11-17]
71. Kawamanto K, Flynn M, Kukhareva P, ElHalta D, Hess R, Gregory T, et al. A pragmatic guide to establishing clinical decision support governance and addressing decision support fatigue: a case study. AMIA Annu Symp Proc 2018;2018:624-633 [FREE Full text] [Medline: 30815104]

Abbreviations

CDSS: clinical decision support system

LBP: low back pain

Edited by G Eysenbach; submitted 26.11.19; peer-reviewed by A Benetoli, L Maclachlan; comments to author 05.01.20; revised version received 06.02.20; accepted 06.02.20; published 11.05.20.

Please cite as:
Downie AS, Hancock M, Abdel Shaheed C, McLachlan AJ, Kocaballi AB, Williams CM, Michaleff ZA, Maher CG
An Electronic Clinical Decision Support System for the Management of Low Back Pain in Community Pharmacy: Development and Mixed Methods Feasibility Study
JMIR Med Inform 2020;8(5):e17203
URL: https://medinform.jmir.org/2020/5/e17203
doi:10.2196/17203
PMID:32390593

Original Paper

The Perceptions of and Factors Associated With the Adoption of the Electronic Health Record Sharing System Among Patients and Physicians: Cross-Sectional Survey

Martin CS Wong[1*], MD, MPH; Junjie Huang[1*], MD, MSc; Paul SF Chan[1], MSc; Veeleah Lok[1], BSc; Colette Leung[1], MSc; Jingxuan Wang[1], MPhil; Clement SK Cheung[2], MD; Wing Nam Wong[2], MD; Ngai Tseung Cheung[2], MD; Chung Ping Ho[3], MRCP, FRCP; Eng Kiong Yeoh[1], MBBS, FRCP, FHKCP

[1]JC School of Public Health and Primary Care, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong

[2]Information Technology and Health Informatics Division, Hospital Authority, Hong Kong

[3]Information Technology Committee, Hong Kong Medical Association, Hong Kong

[*]these authors contributed equally

Corresponding Author:
Martin CS Wong, MD, MPH
JC School of Public Health and Primary Care
Faculty of Medicine
The Chinese University of Hong Kong
4/F, School of Public Health Building, Prince of Wales Hospital
Shatin, NT
Hong Kong, 999077
Phone: 852 2252 8782
Email: wong_martin@cuhk.edu.hk

Abstract

Background: The electronic health record sharing system (eHRSS) was implemented as a new health care delivery platform to facilitate two-way communication between the public and private sectors in Hong Kong.

Objective: This study aimed to investigate the perceptions of and factors associated with the adoption of eHRSS among patients, the general public, and private physicians.

Methods: Telephone interviews were conducted in 2018 by using a simple random sampling strategy from a list of patients who had enrolled in the eHRSS and a territory-wide telephone directory for nonenrolled residents. We completed 2000 surveys (1000 each for enrolled and nonenrolled individuals). Private physicians completed self-administered questionnaires, including 762 valid questionnaires from 454 enrolled physicians and 308 nonenrolled physicians.

Results: Most participants (707/1000, 70.70%) were satisfied with the overall performance of the eHRSS. Regarding registration status, most nonenrolled patients (647/1000, 64.70%) reported that “no recommendation from their physicians and family members” was the major barrier, whereas more than half of the physicians (536/1000, 53.60%) expressed concerns on “additional workload due to use of eHRSS.” A multivariate regression analysis showed that patients were more likely to register when they reported “other service providers could view the medical records” (adjusted odds ratio [aOR] 6.09, 95% CI 4.87-7.63; *P*<.001) and “friends’ or family’s recommendation or assistance in registration” (aOR 3.51, 95% CI 2.04-6.03; *P*=.001). Physicians were more likely to register when they believed that the eHRSS could improve the quality of health care service (aOR 4.70, 95% CI 1.77-12.51; *P*=.002) and were aware that the eHRSS could reduce duplicated tests and treatments (aOR 4.16, 95% CI 1.73-9.97; *P*=.001).

Conclusions: Increasing the possibility of viewing patients' personal medical record, expanding the sharable data scope for patients, and highlighting the benefits of the system for physicians could be effective to enhance the adoption of the eHRSS.

(JMIR Med Inform 2020;8(5):e17452) doi:10.2196/17452

KEYWORDS

electronic health records; hospital shared services; data management

Introduction

Background

Health information technologies, such as electronic health record systems (eHRs), are considered to be critical in transforming health care delivery in terms of improving quality and efficiency [1,2]. In the past decade, eHRs have been launched and implemented in Western countries [3,4]. It was recognized that more extensive adoption of eHRs is effective in reducing medical errors and health care costs, enhancing medical efficacy, and improving health care delivery [5,6]. Nevertheless, the factors associated with the adoption of eHRs remained unknown, especially in Asian regions [7,8].

In Hong Kong, the Public Private Interface-electronic Patient Record (PPI-ePR) program was introduced by the hospital authority (HA) in 2006 as a new electronic platform to enhance data exchange between the public and private sectors [9]. It was the first step toward the vision to develop a territory-wide electronic health record sharing system (eHRSS) that provides a backbone to develop a two-way eHRSS and facilitates better communication between public and private health care services [10]. The eHRSS is a territory-wide health record platform funded by the Food and Health Bureau. The Information and Technology and Health Informatics Department of the HA assisted the government to develop and operate the system as a technical agency. Unlike the mandatory or opt-out enrollment models in similar health record sharing systems in other countries, for example, Denmark, the United Kingdom, or Canada, participation in the eHRSS is opt-in and on a voluntary basis for both patients and health care providers (HCPs).

The eHRSS was launched in March 2016. As of March 2019, over 1,000,000 patients, 47,000 HCPs, all private hospitals (12), and over 1400 HCPs from private sectors, including various types of clinics, elderly homes, and welfare organizations, have enrolled in the eHRSS [11]. With the satisfactory enrollment rates in general, it is an appropriate time to review the current state of the system and areas of enhancement after 3 years of its implementation.

Objectives

This study aimed to investigate the factors associated with registration and adoption of the system among patients and physicians and to examine the awareness, acceptance, perceived benefits, and possible improvements of the eHRSS within the dual health care system of Hong Kong.

Methods

Recruitment

Telephone-based interviews were conducted among enrolled patients and nonenrolled residents. The survey on users was based on a list of enrolled patients provided by the HA, whereas nonusers were selected from the Hong Kong Telephone Directory, which consists of approximately 99% of land-based telephone lines. A simple random sampling methodology was adopted, and computer-generated numbers were used for subject recruitment. We assumed 65.0% as the proportion in all the outcomes. A sample size of approximately 972 enrolled participants will achieve a precision level of 0.03 from the following formula:  , where p=proportion of outcomes. Therefore, we aimed to achieve at least 1000 successful, complete patient surveys each for enrolled and nonenrolled individuals. Assuming a refusal rate of 30%, we made more than 1500 attempts of telephone calls to complete 2000 successful surveys. The response rate was 66.67% (1000/1500) and 60.90% (1000/1642) for enrolled and nonenrolled participants, respectively.

For physician surveys, self-administered questionnaires were conducted among private physicians. Postal addresses of public institutions, nongovernment organizations, or universities were excluded. A list of all enrolled physicians in Hong Kong was provided by the electronic health record office. The response rate of physicians in previous surveys was as low as 5% [12]. To enhance the response rate, one continuous medical education (CME) point was awarded through the Hong Kong Medical Council to each completed physician response. A total of 4340 invitations were sent to private physicians through various channels, including postage, fax lines, email addresses, phone calls, lunchtime seminar programs, and high-concentration buildings where private physicians' practices are located. In total, 762 valid questionnaires were received, consisting of 454 enrolled and 308 nonenrolled physicians. The overall response rate was 17.56% (762/4340).

Survey Instruments

Survey items included enablers and barriers of registration in the eHRSS; the awareness, acceptance, and perceived benefits; reasons for not using the eHRSS after enrollment; and areas for service improvement. The patient and physician surveys were designed by an academic physician with relevant experience in studies related to the eHRSS and extensive expertise in clinical and public health research. The questionnaires drafted were validated by an expert panel of epidemiologists, physicians, nursing professionals, public health practitioners, and academicians. Both surveys were pilot-tested on 20 physicians and 20 patients, respectively, for feasibility and item comprehensiveness. The surveys were available in both Chinese and English versions. All surveys were anonymous. Consent was sought verbally through telephone surveys for patients and by participants' signature through fax or postal surveys for private physicians.

Statistical Analysis

All surveys were checked for completeness and the presence of participant consent. Data entry and analysis were performed using SPSS version 21.0 (IBM Corporation). A random check was conducted to examine the validity, quality, and accuracy of data. A descriptive analysis was performed, and the outcome variables were expressed as proportions. Two binary logistic regression models were constructed for physicians and patients. The first was to examine the predictors of registration (vs no registration), and the second was to evaluate active use (vs inactive use) of the eHRSS after registration. The predictors included (1) sociodemographic factors, (2) perceived usefulness and perceived ease of use based on the technology acceptance

model [13], and (3) cues to action based on variables pertinent to the health belief model [14]. All *P* values ≤.05 were regarded as statistically significant. Variance inflation factors were calculated before the regression analysis. In patients' analysis, 4 variables related to "perceived usefulness" were excluded because of multicollinearity, including "Keep my medical records up-to-date," "Not necessary to bring my medical report," "Reduce my repeated checking and information provision," and "Physicians can get accurate and comprehensive information." Besides, 2 interactions were found to be significant, that is, "souvenirs as an incentive" interacted with "friends' or family's recommendation or assistance in registration" and "the physician's advice or assistance in registration" interacted with "friends' or family's recommendation or assistance in registration." Finally, structural equation modeling (SEM) was adopted to study predictors of patients' registration.

Results

Patient Surveys

Sociodemographic Characteristics

There were more females than males in the enrolled group (426/1000, 42.60% vs 574/1000, 57.40%) and nonenrolled group (332/1000, 33.20% vs 668/1000, 66.80%). Among the enrollees, the majority were aged between 61 and 70 years (291/1000, 29.10%), followed by 71 years or older (282/1000, 28.20%), and between 51 and 60 years (185/1000, 18.50%). Age distributions were similar in nonenrollees, with most aged between 61 and 70 years (234/1000, 23.40%), 71 years or older (229/1000, 22.90%), and between 51 and 60 years (214/1000, 21.40%; Table 1).

Channels of Awareness

Approximately half of the enrolled patients learned about the system from others (487/1000, 48.70%), including hospitals, clinics, health centers, district council members, and social workers. Among them, 31.80% (318/1000) learned about the system from posters or leaflets. Most nonenrollees learned about the eHRSS from television or magazine advertisements (782/1000, 78.20%) and friends and family members (165/1000, 16.50%; Multimedia Appendix 1).

Reasons for No Registration

The majority of nonenrollees (strongly agree or agree: 647/1000, 64.70%) agreed that no recommendations given from their physicians was the major barrier. In addition, approximately half of them (strongly agree or agree: 517/1000, 51.70%) expressed that they only visited one medical professional, and hence, registration was not required. More than one-third of them expressed concerns about the security of personal data and privacy (strongly agree or agree: 480/1000, 48.00%).

Table 1. Sociodemographic characteristics of patients (N=2000).

Variables	Enrollee (n=1000), n (%)	Nonenrollee (n=1000), n (%)	Total, n (%)	*P* value[a]
Gender				<.001
Male	426 (42.6)	332 (33.2)	758 (37.9)	
Female	574 (57.4)	668 (66.8)	1242 (62.1)	
Age (years)				<.001
18-30	27 (2.7)	60 (6.0)	87 (4.4)	
31-40	70 (7.0)	107 (10.7)	177 (8.9)	
41-50	101 (10.1)	150 (15.0)	251 (12.6)	
51-60	185 (18.5)	214 (21.4)	399 (20.0)	
61-70	291 (29.1)	234 (23.4)	525 (26.3)	
≥71	282 (28.2)	229 (22.9)	511 (25.6)	
Refused to answer	44 (4.4)	6 (1.0)	50 (2.5)	
Education				.001
No schooling or preschool education	92 (9.2)	80 (8)	172 (8.6)	
Primary education	270 (27.0)	235 (23.5)	505 (25.3)	
Junior high school	164 (16.4)	135 (13.5)	299 (15.0)	
High school	269 (26.9)	316 (31.6)	585 (29.3)	
Nondegree tertiary education	38 (3.8)	53 (5.3)	91 (4.6)	
Tertiary education	120 (12.0)	155 (15.5)	275 (13.8)	
Others	4 (0)	6 (1.0)	10 (0.5)	
Refused to answer	43 (4.3)	20 (2.0)	63 (3.2)	
Occupation				.002
Full time or part time	292 (29.2)	350 (35.0)	642 (32.1)	
Job-waiting	7 (1.0)	10 (1.0)	17 (0.9)	
Retirement	443 (44.3)	362 (36.2)	805 (40.3)	
Houseworker	206 (20.6)	238 (23.8)	444 (22.2)	
Student	13 (1.3)	17 (1.7)	30 (1.5)	
Others	2 (0)	22 (2.2)	24 (1.2)	
Refused to answer	37 (3.7)	1 (0)	38 (1.9)	
Household income				<.001
<2000	104 (10.4)	94 (9.4)	198 (9.9)	
2000-3999	145 (14.5)	65 (6.5)	210 (10.5)	
4000-5999	67 (6.7)	38 (3.8)	105 (5.3)	
6000-7999	21 (2.1)	23 (2.3)	44 (2.2)	
8000-9999	11 (1.1)	28 (2.8)	39 (2.0)	
10,000-14,999	71 (7.1)	60 (6.0)	131 (7)	
15,000-19,999	50 (5.0)	63 (6.0)	113 (5.7)	
20,000-24,999	73 (7.3)	82 (8.2)	155 (7.8)	
25,000-29,999	35 (3.5)	57 (5.7)	92 (4.6)	
30,000-39,999	54 (5.4)	81 (8.1)	135 (6.8)	
40,000-59,999	43 (4.3)	40 (4.0)	83 (4.2)	
≥60,000	48 (4.8)	25 (2.5)	73 (3.7)	
Refused to answer	278 (27.8)	344 (34.4)	622 (31.1)	

Variables	Enrollee (n=1000), n (%)	Nonenrollee (n=1000), n (%)	Total, n (%)	*P* value[a]
Joined the public private interface electronic patient record program				<.001
Yes	20 (2.0)	9 (1.0)	29 (1.5)	
No	952 (95.2)	985 (98.5)	1937 (96.9)	
Refused to answer	28 (2.8)	6 (1.0)	34 (1.7)	
Required regular follow-up consultation				<.001
Yes	602 (60.2)	274 (27.4)	876 (43.8)	
No	390 (39.0)	719 (71.9)	1109 (55.5)	
Refused to answer	8 (1.0)	7 (1.0)	15 (0.8)	

[a]Proportions were compared by using chi-square tests.

Reasons for Not Using the System After Registration

For the enrollees who did not use the system (498 out of 1000), the reasons they did not do so after registration were “they were not sick after participation” (strongly agree or agree: 221/498, 45.5%), “they only went to one place to see a physician” (strongly agree or agree: 240/498, 49.4%), and “they did not tell the physician that they had registered (strongly agree or agree: 115/498, 23.8%).

Level of Satisfaction Among the Patients

Most enrollees were satisfied with the eHRSS, with 70.70% (707/1000) of the enrollees reporting that they were satisfied or strongly satisfied. Regarding the registration process, 91.20% (912/1000) of the enrollees reported that they were satisfied or strongly satisfied with the registration procedures and registration methods.

Perceived Areas for Future Improvement

Most of the enrollees suggested that they should be able to access their medical records through the system (30/124, 24.2%) and more sharable information (32/124, 25.8%). Others recommended that the system should be designed in a more comprehensive and user-friendly manner (23/124, 18.6%), involve the participation of more physicians (16/124, 12.9%), and increase publicity (10/124, 8.1%; Multimedia Appendix 2).

Factors Associated With Registration and Usage

Regarding the status of registration (Table 2), patients were more likely to register when they (1) were in the highest household income group (HK $60,000 [US $7696] or above; reference: income <14,999 [US $1924]; aOR 2.28, 95% CI 1.17-4.46; *P*=.02), (2) needed regular clinic follow-up (aOR 3.49, 95% CI 2.70-4.50; *P*<.001), (3) reported “other service providers could view the medical records” (aOR 6.09, 95% CI 4.87-7.63; *P*<.001) as perceived usefulness of the eHRSS, and (4) reported “friends’ or family’s recommendation or assistance in registration” (aOR 3.51, 95% CI 2.04-6.03; *P*=.001) as one of the cues to action.

Table 2. Factors associated with the status of registration and usage of the system among patients.

Variables	Status of registration		Usage of the system	
	Adjusted odds ratio (aOR; 95% CI)	*P* value	aOR (95% CI)	*P* value
Gender				
Male	1 (Ref[a])	N/A[b]	1 (Ref)	N/A
Female	0.71 (0.55-0.91)	.008	1.18 (0.87-1.59)	.29
Age (years)				
18-40	1 (Ref)	N/A	1 (Ref)	N/A
41-60	1.00 (0.66-1.51)	.99	1.20 (0.82-3.27)	.16
≥61	1.28 (0.79-2.06)	.32	1.73 (0.93-3.19)	.08
Education				
Primary or below	1 (Ref)	N/A	1 (Ref)	N/A
Secondary	0.96 (0.72-1.29)	.79	0.98 (0.69-1.37)	.89
Tertiary or above	0.94 (0.62-1.43)	.79	1.22 (0.73-2.02)	.45
Occupation				
Working (full time or part time)	1 (Ref)	N/A	1 (Ref)	N/A
Not working (searching for a job, retired, houseworker, or student)	0.72 (0.52-1.01)	.06	1.05 (0.70-1.56)	.81
Household income (HK $)				
≤14,999 (US $1924)	1 (Ref)	N/A	1 (Ref)	N/A
15,000-24,999 (US $3207)	0.67 (0.47-0.97)	.03	0.70 (0.45-1.08)	.11
25,000-59,999 (US $7696)	0.61 (0.38-0.97)	.04	0.68 (0.39-1.19)	.17
≥60,000 (US $7696)	2.28 (1.17-4.46)	.02	0.46 (0.21-0.97)	.04
Refused to answer	0.81 (0.59-1.10)	.18	0.84 (0.57-1.22)	.36
Joined the public private interface electronic patient record program				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	2.01 (0.72-5.57)	.18	1.46 (0.56-3.81)	.44
Required regular follow-up				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	3.49 (2.70-4.50)	<.001	1.65 (1.20-2.26)	.002
Perceived usefulness				
Other medical service providers can view the medical records	6.09 (4.87-7.63)	<.001	1.71 (1.31-2.23)	<.001
Cues to action				
Souvenir	1.66 (0.97-2.84)	.07	4.80 (2.72-8.48)	<.001
Friends' or family's recommendation or assistance in registration	3.51 (2.04-6.03)	.001	2.07 (1.43-2.98)	<.001
Doctor's advice or assistance in registration	1.25 (0.85-1.84)	.27	1.52 (1.14-2.02)	.004
Interaction effects				
Interaction 1[c]	0.77 (0.66-0.89)	<.001	0.64 (0.54-0.77)	<.001
Interaction 2[d]	0.72 (0.63-0.81)	<.001	0.91 (0.84-1.00)	.04

[a]Ref: reference group in the regression analysis.

[b]N/A: not applicable.

[c]Souvenirs and friends' or family's recommendation or assistance in registration.

[d]Doctor's advice or assistance in registration and friends' or family's recommendation or assistance in registration.

Regarding the usage of the system (Table 2), enrollees were more likely to use the system when they (1) needed regular follow-up (aOR 1.65, 95% CI 1.20-2.26; *P*=.002), (2) reported that other service providers could view the medical records (aOR 1.71, 95% CI 1.31-2.23; *P*<.001), (3) reported physicians' advice or assistance in registration (aOR 1.52, 95% CI 1.14-2.02; *P*=.004), (4) reported friends' or family's recommendation or assistance in registration (aOR 2.07, 95% CI 1.43-2.98; *P*<.001), and (5) were provided with souvenirs (aOR 4.80, 95% CI 2.72-8.48; *P*<.001). The effect size of the souvenir is among the largest, followed by friends' or family's recommendation and the needs of regular follow-up.

SEM was adopted to study the predictors of patients' registration (Multimedia Appendix 3). In this model, associations of observed variables to the latent variables were strong. The 2 observed variables, "friends' or family's recommendation or assistance" and "the physician's advice or assistance," had factor loadings of 0.79 and 0.66, respectively, with cues to action (latent variable). The other 4 observed variables, "reduce my repeated checking and information provision," "keep my medical records up-to-date," "doctors can get accurate and comprehensive information," and "other HCPs can read my medical records," had factor loadings between 0.93 and 0.99 with perceived benefits (latent variable). Cues to actions influenced perceived benefits with a magnitude of 0.35, and perceived benefits determined the status of registration with a magnitude of 0.38.

Physician Surveys

Sociodemographic Characteristics

There were more male than female participants among the enrollees (314/454, 69.2% vs 105/454, 23.1%) and nonenrollees (216/308, 70.1% vs 64/308, 20.8%). In general, the enrollees (271/454, 59.7%; aged between 41 and 60 years) were younger than the nonenrollees (127/308, 41.2%; aged 61 years or older; Table 3).

Table 3. Sociodemographic characteristics of physicians (N=762).

Variables	Enrollee (n=454), n (%)	Nonenrollee (n=308), n (%)	Total, n (%)	*P* value[a]
Gender				.51
Male	314 (69.2)	216 (70.1)	530 (69.6)	
Female	105 (23.1)	64 (20.8)	169 (22.2)	
Missing	35 (7.7)	28 (9.1)	63 (8.3)	
Age (years)				<.001
≤30	3 (0.7)	0 (0)	3 (0.4)	
31-40	47 (10.4)	14 (4.5)	61 (8.0)	
41-50	122 (26.9)	61 (19.8)	183 (24.0)	
51-60	149 (32.8)	86 (27.9)	235 (30.8)	
≥61	102 (22.5)	127 (41.2)	229 (30.1)	
Missing	31 (6.8)	20 (6.5)	51 (6.7)	
Years of practice				<.001
≤4	4 (0.9)	0 (0)	4 (0.5)	
5-9	12 (2.6)	3 (1.0)	15 (2.0)	
10-19	107 (23.6)	50 (16.2)	157 (20.6)	
20-29	120 (26.4)	62 (20.1)	182 (23.9)	
≥30	178 (39.2)	170 (55.2)	348 (45.7)	
Missing	33 (7.3)	23 (7.5)	56 (7.3)	
Type of institution				<.001
Solo practice	223 (49.1)	180 (58.4)	403 (52.9)	
With partners or group practice	150 (33.0)	70 (22.7)	220 (28.9)	
Private hospital	31 (6.8)	13 (4.2)	44 (5.8)	
Others	19 (4.2)	14 (4.5)	33 (4.3)	
Missing	31 (6.8)	31 (10.1)	62 (8.1)	
Specialty				.006
Nil	108 (23.8)	95 (30.8)	203 (26.6)	
Anesthesiology	1 (0.2)	2 (0.6)	3 (0.4)	
Community medicine	3 (0.7)	2 (0.6)	5 (0.7)	
Emergency medicine	3 (0.7)	1 (0.3)	4 (0.5)	
Family medicine	79 (17.4)	29 (9.4)	108 (14.2)	
Internal medicine	61 (13.4)	18 (5.8)	79 (10.4)	
Obstetrics and gynecology	24 (5.3)	23 (7.5)	47 (6.2)	
Ophthalmology	13 (2.9)	7 (2.3)	20 (2.6)	
Orthopedics and traumatology	20 (4.4)	9 (2.9)	29 (3.8)	
Otorhinolaryngology	8 (1.8)	6 (1.9)	14 (1.8)	
Pediatrics	24 (5.3)	21 (6.8)	45 (5.9)	
Pathology	1 (0.2)	2 (0.6)	3 (0.4)	
Psychiatry	8 (1.8)	24 (7.8)	32 (4.2)	
Radiology	4 (0.9)	7 (2.3)	11 (1.4)	
Surgery	48 (10.6)	16 (5.2)	64 (8.4)	
Others	35 (7.7)	27 (8.8)	62 (8.1)	
Missing	35 (7.7)	25 (8.1)	60 (7.9)	

Variables	Enrollee (n=454), n (%)	Nonenrollee (n=308), n (%)	Total, n (%)	*P* value[a]
Joined the public private interface electronic patient record program				<.001
Yes	355 (78.2)	50 (16.2)	405 (53.1)	
No	85 (18.7)	251 (81.5)	336 (44.1)	
Missing	14 (3.1)	7 (2.3)	21 (2.8)	

[a]Proportions were compared by using chi-square tests.

Channels of Awareness

Approximately 39.4% (179/454) of the enrollees were aware of the system from peers in the health care sector, followed by practice clinics (174/454, 38.3%) and government-subsidized programs (119/454, 26.2%). For the 284 nonenrolled physicians who were aware of the system, the modes of receiving the information were as follows: mainly from peers in the health care sector (136/284, 47.9%), television or magazine advertisements (92/284, 32.4%), and posters or website (87/284, 30.6%; Multimedia Appendix 4).

Reasons for No Registration

More than half of the participants expressed concerns about the additional workload (strongly agree or agree: 166/308, 53.6%), whereas 45.5% (140/308) perceived the enrollment procedures to be complicated.

Reasons for Not Using the System After Registration

In addition, 6.8% (31/454) of enrollees did not access any patients' medical record after the registration. Among them, 42% (13/31) stated that there was no clinical indication for accessing the data, followed by technical issues such as forgetting the log-in password (6/31, 19%) and "patients not using the system" or "patients did not inform their registration status" (6/31, 19%).

Level of Satisfaction Among the Physicians

Most enrollees were satisfied with the system, with 50.2% (228/454) and 7.7% (35/454) of the enrollees reporting being "satisfied" and "strongly satisfied," respectively. A similar level of satisfaction was observed for "Instructions for use" (satisfied: 200/454, 44.1%; strongly satisfied: 35/454, 7.7%) and "compatibility of Web browser" (satisfied: 196/454, 43.2%; strongly satisfied: 34/454, 7.5%).

Perceived Areas for Future Improvement

Simplification of the enrollment process (enrollees: 190/454, 41.9%; nonenrollees: 166/308, 53.9%), provision of technical support (enrollees: 157/454, 34.6%; nonenrollees: 161/308, 52.3%), and improvement of interface friendliness (enrollees: 197/454, 43.4%; nonenrollees: 136/308, 44.2%) were the most commonly chosen options among physicians. Notably, over half of the enrollees (268/454, 59.0%) suggested to expand the sharable data scope (Multimedia Appendix 5), and the radiology image was the most commonly chosen option (enrollees: 335/454, 73.8%; nonenrollees: 231/308, 75%; Multimedia Appendix 6).

Perceived Strategies to Increase the Awareness

Traditional channels such as "television or newspaper or magazine advertisement" (enrollees: 259/454, 57.1%; nonenrollees: 112/308, 57.1%), academic publications such as medical newsletters and journals (enrollees: 168/454, 37%; nonenrollees: 161/308, 52.3%), and new media including website or social media (enrollees: 194/454, 42.7%; nonenrollees: 112/308, 36.4%) were perceived as effective strategies among the physician participants (Multimedia Appendix 7).

Factors Associated With Registration and Usage

Physicians were more likely to register for the eHRSS when they (1) had previously joined PPI-ePR (aOR 69.20, 95% CI 31.41-152.45; *P*<.001), (2) believed that it could improve the quality of health care service (aOR 4.70, 95% CI 1.77-12.51; *P*=.002), or (3) were aware that it could reduce duplicated tests and treatments (aOR 4.16, 95% CI 1.73-9.97; *P*=.001; Table 4).

Table 4. Factors associated with the status of registration among physicians.

Variables	Crude odds ratio (95% CI)	*P* value	Adjusted odds ratio (95% CI)	*P* value
Gender				
Male	1 (Ref[a])	N/A[b]	1 (Ref)	N/A
Female	1.13 (0.79-1.61)	.51	1.15 (0.54-2.42)	.72
Age (years)				
≤40	1 (Ref)	N/A	1 (Ref)	N/A
41-60	0.52 (0.28-0.97)	.04	0.41 (0.12-1.37)	.15
≥61	0.22 (0.12-0.43)	<.001	0.25 (0.06-1.09)	.06
Types of medical practice				
Solo	1 (Ref)	N/A	1 (Ref)	N/A
With partner or group	1.73 (1.22-2.44)	.002	1.13 (0.57-2.25)	.73
Private hospital	1.92 (0.98-3.79)	.06	2.54 (0.60-10.81)	.21
Others	1.10 (0.53-2.25)	.80	2.18 (0.52-9.20)	.29
Years of practice				
≤9	1 (Ref)	N/A	1 (Ref)	N/A
10-29	0.38 (0.11-1.33)	.13	1.03 (0.09-11.28)	.98
≥30	0.20 (0.06-0.69)	.01	0.69 (0.06-8.28)	.77
Joined the public private interface electronic patient record program				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	20.97 (14.27-30.81)	<.001	69.20 (31.41-152.45)	<.001
Perceived ease of use				
Timely access				
Disagree or strongly disagree	1 (Ref)	N/A	1 (Ref)	N/A
Neutral	1.51 (0.80-2.84)	.20	2.03 (0.64-6.46)	.23
Agree or strongly agree	3.48 (2.00-6.05)	<.001	2.67 (0.97-7.34)	.06
Not applicable	0.27 (0.08-0.90)	.03	0.03 (0.01-0.24)	.001
Cues to action				
As required by subsidized program				
Disagree or strongly disagree	1 (Ref)	N/A	1 (Ref)	N/A
Neutral	0.53 (0.32-0.89)	.02	0.31 (0.12-0.86)	.02
Agree or strongly agree	0.50 (0.31-0.82)	.006	0.49 (0.20-1.20)	.12
Not applicable	1.69 (0.95-2.99)	.07	1.72 (0.56-5.25)	.34
Perceived benefits				
Quality improvement				
No	1 (Ref)	N/A	1 (Ref)	N/A
Maybe	0.79 (0.42-1.50)	.48	1.36 (0.52-3.60)	.53
Yes	5.10 (2.75-9.44)	<.001	4.70 (1.77-12.51)	.002
Comprehensiveness				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	2.19 (1.59-3.01)	<.001	0.70 (0.32-1.53)	.37
Reduction of errors				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	1.97 (1.44-2.70)	<.001	0.57 (0.25-1.30)	.18

Variables	Crude odds ratio (95% CI)	*P* value	Adjusted odds ratio (95% CI)	*P* value
Reduction of duplicates				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	3.87 (2.72-5.51)	<.001	4.16 (1.73-9.97)	.001
Accuracy and timely access				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	2.52 (1.85-3.44)	<.001	1.77 (0.79-3.94)	.16
Disease surveillance and monitoring				
No	1 (Ref)	N/A	1 (Ref)	N/A
Yes	1.48 (1.06-2.08)	.02	1.17 (0.52-2.63)	.71

[a]Ref: reference group in the regression analysis.

[b]N/A: not applicable.

Regarding the usage of the system (Table 5), insignificant results were observed for all variables in the multivariate logistic regression model. Therefore, a univariate analysis was performed to study their likelihood to use the system. Variables were reported when their *P* values were ≤.20. From Table 5, we can observe that physicians are more likely to use the system when they (1) have previously joined PPI-ePR (crude odds ratio [COR] 6.58, 95% CI 3.05-14.17; *P*<.001), (2) agreed that meeting patients' request was a reason for enrolling in the eHRSS (COR 3.21, 95% CI 1.05-9.84; *P*=.04), (3) believed that the system could improve the quality of health care service (COR 5.34, 95% CI 1.35-21.04; *P*=.02), or (4) were aware that the system could reduce duplicated tests and treatments (COR 3.11, 95% CI 1.38-7.04; *P*<.006).

Table 5. Factors associated with the usage of system among physicians.

Variables	Crude odds ratio (95% CI)	*P* value
Gender		
Male	1 (Ref[a])	N/A[b]
Female	0.52 (0.24-1.14)	.10
Joined the public private interface electronic patient record program		
No	1 (Ref)	N/A
Yes	6.58 (3.05-14.17)	<.001
Perceived ease of use		
Timely access		
Disagree or strongly disagree	1 (Ref)	N/A
Neutral	1.25 (0.29-5.44)	.77
Agree or strongly agree	3.57 (0.95-13.45)	.06
Instruction of use		
Disagree or strongly disagree	1 (Ref)	N/A
Neutral	0.44 (0.12-1.59)	.21
Agree or strongly agree	2.75 (0.60-12.61)	.19
Compatibility of web browser		
Disagree or strongly disagree	1 (Ref)	N/A
Neutral	0.65 (0.22-1.94)	.44
Agree or strongly agree	2.88 (0.81-10.23)	.10
Cues to action		
Patients' request	1 (Ref)	N/A
Disagree or strongly disagree	1.50 (0.45-4.98)	.51
Neutral	3.21 (1.05-9.84)	.04
Agree or strongly agree	0.31 (0.06-1.56)	.16
Perceived benefits		
Quality improvement		
No	1 (Ref)	N/A
Yes	5.34 (1.35-21.04)	.02
Reduction of duplicates		
No	1 (Ref)	N/A
Yes	3.11 (1.38-7.04)	.006
Disease surveillance and monitoring		
No	1 (Ref)	N/A
Yes	0.58 (0.27-1.24)	.16

[a]Ref: reference group in the regression analysis.

[b]N/A: not applicable.

Discussion

Principal Findings

Overall, both patients and physicians were satisfied with the eHRSS. Nonenrolled patients were aware of the system mainly from traditional communication channels (television or magazine advertisements), whereas nearly half of the enrolled patients learned about it via hospitals or clinics, community centers, district council members, and social workers. Physicians learned about the eHRSS from their peers in the health care sector. The most important factor hindering system enrollment of nonenrolled patients was the absence of recommendations from their physicians. In addition, they only visited one medical physician, and hence, registration in the system was not needed.

Nonenrolled physicians were concerned about the potential increase of workload after registration and perceived the enrollment procedure as complicated. Patients did not use the system after registration mostly because they had no such need or opportunity, whereas enrolled physicians did not utilize the system as they did not perceive any clinical indication for data access.

Explanation of Findings and Comparison With Prior Work

The survey findings reported the factors that hindered enrollment among patients. The most significant factor was the absence of recommendations from their physicians. Previous literature has demonstrated that people who appear to have authority can help a person make a particular decision [15]. Other factors included concerns about personal data and privacy issues, and the uncertainty about benefits of the system. The main reasons for not registering among physicians included perceived additional workload and complicated enrollment procedures. Evidence showed perceived workload and ease of use for a system was positively associated with its adoption [16,17]. Physicians were more likely to register when they thought that the system would improve health care quality and reduce the duplication of work, which was consistent with our findings [18]. Two reasons for not using the system among the enrolled patients were that there was no registration among physicians and that they did not inform the physicians that they had registered. As for those enrolled physicians who did not access the medical records via the system after registration, technical issues such as forgetting the log-in password were among the major reasons, and this observation is consistent with previous results [17].

Some studies have been performed on eHRs in Western countries, including the United Kingdom, the United States, and Canada [19,20]. For instance, in England, a multilevel case study with 216 participants consisting of patients, clinical staff, and project managers was conducted to investigate the use of personal electronic health records in 2010 [21]. The results showed that most of the participants perceived it neither useful nor easy to use. Nevertheless, the researchers in this case study also acknowledged that these findings should be interpreted with caution given the small sample size. As for the United States, there were 54% of physicians who adopted the eHRs in 2011 [22]. Most of the physicians who adopted an eHRs reported being satisfied with their system. Approximately half of the users agreed that the system could improve patient care. Perceived management support, provider involvement, and adequate training were the main facilitators, whereas perceived lack of usefulness and provider autonomy were the major barriers in its adoption [16]. A cross-sectional study among Canadian medical practitioners, involving 102 users and 83 nonusers, found that perceived ease of use was the strongest facilitator for eHRs use, whereas usefulness and ease of use were the main factors influencing system adoption among nonusers [23]. Although Asian countries or regions such as Japan, Taiwan, and Singapore have initiated the development of eHRs, there was a lack of studies on perceptions, awareness, and factors of adoption of the system [24].

Strengths and Limitations

This study has comprehensively evaluated the perceptions, acceptance, and factors of eHRSS adoption, which has been implemented in Hong Kong since 2016. Although the benefits, facilitators, and barriers of eHRs have been widely discussed in Western countries, including the United Kingdom [21], Canada [23], and the United States [22], in the past decade, much effort is needed in Asian cities where eHRs were generally established in the past few years [24,25]. Meanwhile, previous studies mainly focused on either patients' or physicians' perspectives [26,27]. Our study included perceptions among enrolled patients, nonenrolled patients, enrolled physicians, and nonenrolled physicians. In addition, an updated patient list that contained enrolled patients and a territory-wide telephone directory for nonenrolled patients were used with a simple random sampling strategy, which enhanced the generalizability of our findings.

There are several limitations of this study. First, the survey was a cross-sectional study and could not establish a cause-and-effect relationship because of the possibility of reverse causality. Prospective longitudinal studies are required to confirm the facilitators and barriers. In addition, the survey questions were designed through face validity rather than construct validity. The consistency reliability of the survey measurements was yet to be evaluated. In addition, the overall response rate among physician participants was low (17.6%), and it might have caused nonresponse bias. However, the study adopted different strategies to enhance the response rate, including CME point, postage, fax lines, email addresses, phone calls, lunchtime seminar programs, and visits to high-concentration buildings where private physicians' practices are located. Hence, the response rate was much higher than that in the previous local study (5%). Finally, there may be other variables that could affect the registration and adoption of eHRSS, and hence, some residual confounders may remain uncontrolled.

Lessons Learned

Findings of this study can inform future clinical practice and public health policy on the promotion of eHRSS adoption. To enhance the enrollment rate of eHRSS among patients who have not yet registered, recommendations by primary care physicians during their daily clinical practice is considered to be the most influential factor. It is also important to deliver a sense of adequate and appropriate security protection to the public because it is another key concern for the adoption of eHRSS among patients [28,29]. Multilevel measurements are needed to protect personal data in the eHRSS, such as consent-based record sharing, role-based access control, full data encryption, as well as network and application security defense and protection. In addition, the awareness of the benefits of the eHRSS should be enhanced in the community. To achieve this, future promotional campaigns and educational seminars on the benefits of eHRSS can be effective based on findings from previous evaluations [16]. As for the primary care physicians, communication among the physician users may influence the use of eHRSS. The study found that physician users learned about the system most commonly from their peers in the health care sectors. Therefore, more interviews of the enrolled

physicians in electronic health (eHealth) news and booths in physicians' conferences could be organized to promote the adoption of eHRSS among them [30]. To enhance the actual use of the eHRSS after enrollment among patients, efforts to improve the enrollment among physicians can be effective as it was found to be the most significant factor associated with its use. For physicians who have already enrolled in the eHRSS, it is suggested to provide easier channels for them to retrieve passwords in case they were forgotten. In addition, more technical support on the system could be provided and the user-friendliness of the system interface could be enhanced to maintain long-term adoption of the eHRSS by reducing the time spent on dealing with technical issues.

Conclusions

Participants were satisfied with the overall performance of the system. For patients, the possibility of viewing their personal medical records and expanding the sharable data scope in the system could be a future direction of development. In addition, messages about the stringent measures in protecting privacy and benefits of the system should be clearly conveyed to the public. For physicians, major barriers of registration and usage, such as perceived additional workload, complicated procedures, and lack of technical assistance, will require additional practical and logistic support. It is recommended to enlist enrolled physicians to promote the system among their peer colleagues, such as more interviews in eHealth news and booths in physicians' conferences.

Acknowledgments

This project was funded by the HA of Hong Kong Special Administrative Region (project number: 8110-059-888). The funders had no role in the study design, data collection, data analysis, data interpretation, report writing, or decision to submit the manuscript for publication. The corresponding author had full access to all the data in the study and had final responsibility in the decision to submit for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Channels for patients to know about electronic health record sharing system.
[PDF File (Adobe PDF File), 57 KB - medinform_v8i5e17452_app1.pdf]

Multimedia Appendix 2
Perceived areas to improve electronic health record sharing system among patients.
[PDF File (Adobe PDF File), 56 KB - medinform_v8i5e17452_app2.pdf]

Multimedia Appendix 3
Factors associated with electronic health record sharing system registration among patients: structural equation modeling .
[PDF File (Adobe PDF File), 66 KB - medinform_v8i5e17452_app3.pdf]

Multimedia Appendix 4
Channels for physicians to know about electronic health record sharing system.
[PDF File (Adobe PDF File), 331 KB - medinform_v8i5e17452_app4.pdf]

Multimedia Appendix 5
Perceived areas to improve electronic health record sharing system among physicians.
[PDF File (Adobe PDF File), 59 KB - medinform_v8i5e17452_app5.pdf]

Multimedia Appendix 6
Perceived scope of areas to be expanded among physicians.
[PDF File (Adobe PDF File), 52 KB - medinform_v8i5e17452_app6.pdf]

Multimedia Appendix 7
Perceived strategies to increase the awareness of electronic health record sharing system among physicians.
[PDF File (Adobe PDF File), 60 KB - medinform_v8i5e17452_app7.pdf]

References

1. Buntin MB, Burke MF, Hoaglin MC, Blumenthal D. The benefits of health information technology: a review of the recent literature shows predominantly positive results. Health Aff (Millwood) 2011 Mar;30(3):464-471. [doi: 10.1377/hlthaff.2011.0178] [Medline: 21383365]
2. Shekelle PG, Morton SC, Keeler EB. Costs and benefits of health information technology. Evid Rep Technol Assess (Full Rep) 2006 Apr(132):1-71. [doi: 10.23970/ahrqepcerta132] [Medline: 17627328]
3. Lawrence JE, Cundall-Curry D, Stewart ME, Fountain DM, Gooding CR. The use of an electronic health record system reduces errors in the National Hip Fracture Database. Age Ageing 2019 Mar 1;48(2):285-290. [doi: 10.1093/ageing/afy177] [Medline: 30395143]
4. Fields D, Riesenmy K, Blum TC, Roman PM. Implementation of electronic health records and entrepreneurial strategic orientation in substance use disorder treatment organizations. J Stud Alcohol Drugs 2015 Nov;76(6):942-951 [FREE Full text] [doi: 10.15288/jsad.2015.76.942] [Medline: 26562603]
5. Rizvi RF, Marquard JL, Hultman GM, Adam TJ, Harder KA, Melton GB. Usability evaluation of electronic health record system around clinical notes usage-an ethnographic study. Appl Clin Inform 2017 Oct;8(4):1095-1105 [FREE Full text] [doi: 10.4338/ACI-2017-04-RA-0067] [Medline: 29241247]
6. Howley MJ, Chou EY, Hansen N, Dalrymple PW. The long-term financial impact of electronic health record implementation. J Am Med Inform Assoc 2015 Mar;22(2):443-452. [doi: 10.1136/amiajnl-2014-002686] [Medline: 25164255]
7. Singh K, Johnson L, Devarajan R, Shivashankar R, Sharma P, Kondal D, et al. Acceptability of a decision-support electronic health record system and its impact on diabetes care goals in South Asia: a mixed-methods evaluation of the CARRS trial. Diabet Med 2018 Dec;35(12):1644-1654. [doi: 10.1111/dme.13804] [Medline: 30142228]
8. Inokuchi R, Sato H, Nakamura K, Aoki Y, Shinohara K, Gunshin M, et al. Motivations and barriers to implementing electronic health records and ED information systems in Japan. Am J Emerg Med 2014 Jul;32(7):725-730. [doi: 10.1016/j.ajem.2014.03.035] [Medline: 24792932]
9. Hospital Authority of Hong Kong Government. What is PPI-ePR? URL: https://tinyurl.com/ycc4gfw5 [accessed 2020-01-22]
10. Cheung NT, Fung V, Wong WN, Tong A, Sek A, Greyling A, et al. Principles-based medical informatics for success--how Hong Kong built one of the world's largest integrated longitudinal electronic patient records. Stud Health Technol Inform 2007;129(Pt 1):307-310. [Medline: 17911728]
11. The Government of the Hong Kong Special Administrative Region. Electronic Health Record Sharing System Records Millionth Patient Registrant at Third Anniversary URL: https://wwwinfogovhk/gia/general/201903/05/P20190305003 [accessed 2020-01-22]
12. Cheung CS, Tong EL, Cheung NT, Chan WM, Wang HH, Kwan MW, et al. Factors associated with adoption of the electronic health record system among primary care physicians. JMIR Med Inform 2013 Aug 26;1(1):e1 [FREE Full text] [doi: 10.2196/medinform.2766] [Medline: 25599989]
13. Nunes A, Limpo T, Castro SL. Acceptance of mobile health applications: examining key determinants and moderators. Front Psychol 2019;10:2791 [FREE Full text] [doi: 10.3389/fpsyg.2019.02791] [Medline: 31920836]
14. Ahadzadeh AS, Sharif SP, Ong FS, Khong KW. Integrating health belief model and technology acceptance model: an investigation of health-related internet use. J Med Internet Res 2015 Feb 19;17(2):e45 [FREE Full text] [doi: 10.2196/jmir.3564] [Medline: 25700481]
15. Cialdini RB. Influence: Science And Practice. Fourth Edition. Boston: Allyn & Bacon; 2001.
16. Hamid F, Cline TW. Providers' acceptance factors and their perceived barriers to Electronic Health Record (EHR) Adoption. J Nurs Inform 2013;17(3):1-11 [FREE Full text]
17. Soares N, Vyas K, Perry B. Clinician perceptions of pediatric growth chart use and electronic health records in Kentucky. Appl Clin Inform 2012;3(4):437-447 [FREE Full text] [doi: 10.4338/ACI-2012-06-RA-0023] [Medline: 23646089]
18. Hudson JS, Neff JA, Padilla MA, Zhang Q, Mercer LT. Predictors of physician use of inpatient electronic health records. Am J Manag Care 2012 Apr;18(4):201-206 [FREE Full text] [Medline: 22554008]
19. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption factors of the electronic health record: a systematic review. JMIR Med Inform 2016 Jun 1;4(2):e19 [FREE Full text] [doi: 10.2196/medinform.5525] [Medline: 27251559]
20. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. J Med Syst 2016 Dec;40(12):252 [FREE Full text] [doi: 10.1007/s10916-016-0628-9] [Medline: 27714560]
21. 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]
22. Jamoom E, Beatty P, Bercovitz A, Woodwell D, Palso K, Rechtsteiner E. Physician adoption of electronic health record systems: United States, 2011. NCHS Data Brief 2012 Jul(98):1-8 [FREE Full text] [Medline: 23050588]
23. Archer N, Cocosila M. A comparison of physician pre-adoption and adoption views on electronic health records in Canadian medical practices. J Med Internet Res 2011 Aug 12;13(3):e57 [FREE Full text] [doi: 10.2196/jmir.1726] [Medline: 21840835]
24. Ghani MK, Bali RK, Naguib RN, Marshall IM, Nilmini SW. Electronic health records approaches and challenges: a comparison between Malaysia and four East Asian countries. Int J Electron Healthc 2008;4(1):78-104. [doi: 10.1504/IJEH.2008.018922] [Medline: 18583297]

25. Sittig DF. Personal health records on the internet: a snapshot of the pioneers at the end of the 20th Century. Int J Med Inform 2002 Apr;65(1):1-6. [doi: 10.1016/s1386-5056(01)00215-5] [Medline: 11904243]
26. Wiljer D, Urowitz S, Apatu E, DeLenardo C, Eysenbach G, Harth T, Canadian Committee for Patient Accessible Health Records. Patient accessible electronic health records: exploring recommendations for successful implementation strategies. J Med Internet Res 2008 Oct 31;10(4):e34 [FREE Full text] [doi: 10.2196/jmir.1061] [Medline: 18974036]
27. Miller DP, Latulipe C, Melius KA, Quandt SA, Arcury TA. Primary care providers' views of patient portals: interview study of perceived benefits and consequences. J Med Internet Res 2016 Jan 15;18(1):e8 [FREE Full text] [doi: 10.2196/jmir.4953] [Medline: 26772771]
28. Kruse CS, Mileski M, Alaytsev V, Carol E, Williams A. Adoption factors associated with electronic health record among long-term care facilities: a systematic review. BMJ Open 2015 Jan 28;5(1):e006615 [FREE Full text] [doi: 10.1136/bmjopen-2014-006615] [Medline: 25631311]
29. Ben-Zion R, Pliskin N, Fink L. Critical success factors for adoption of electronic health record systems: literature review and prescriptive analysis. Inf Syst Manag 2014;31(4):296-312. [doi: 10.1080/10580530.2014.958024]
30. Kruse CS, DeShazo J, Kim F, Fulton L. Factors associated with adoption of health information technology: a conceptual model based on a systematic review. JMIR Med Inform 2014 May 23;2(1):e9 [FREE Full text] [doi: 10.2196/medinform.3106] [Medline: 25599673]

Abbreviations

aOR: adjusted odds ratio
CME: continuous medical education
COR: crude odds ratio
eHealth: electronic health
eHRs: electronic health record systems
eHRSS: electronic health record sharing system
HA: hospital authority
HCP: health care provider
PPI-ePR: Public Private Interface-electronic Patient Record
SEM: structural equation modeling

Edited by C Lovis; submitted 13.12.19; this is a non–peer-reviewed article;accepted 09.02.20; published 21.05.20.

Please cite as:
Wong MCS, Huang J, Chan PSF, Lok V, Leung C, Wang J, Cheung CSK, Wong WN, Cheung NT, Ho CP, Yeoh EK
The Perceptions of and Factors Associated With the Adoption of the Electronic Health Record Sharing System Among Patients and Physicians: Cross-Sectional Survey
JMIR Med Inform 2020;8(5):e17452
URL: http://medinform.jmir.org/2020/5/e17452/
doi:10.2196/17452
PMID:32436855

Original Paper

Categorization of Third-Party Apps in Electronic Health Record App Marketplaces: Systematic Search and Analysis

Jordon Ritchie[1*], BSc; Brandon Welch[1*], MSc, PhD

Medical University of South Carolina, Charleston, SC, United States

[*]all authors contributed equally

Corresponding Author:
Brandon Welch, MSc, PhD
Medical University of South Carolina
135 Cannon Street, MSC 200
Suite 405
Charleston, SC, 29425
United States
Phone: 1 8435183769
Email: welchbm@musc.edu

Abstract

Background: Third-party electronic health record (EHR) apps allow health care organizations to extend the capabilities and features of their EHR system. Given the widespread utilization of EHRs and the emergence of third-party apps in EHR marketplaces, it has become necessary to conduct a systematic review and analysis of apps in EHR app marketplaces.

Objective: The goal of this review is to organize, categorize, and characterize the availability of third-party apps in EHR marketplaces.

Methods: Two informaticists (authors JR and BW) used grounded theory principles to review and categorize EHR apps listed in top EHR vendors' public-facing marketplaces.

Results: We categorized a total of 471 EHR apps into a taxonomy consisting of 3 primary categories, 15 secondary categories, and 55 tertiary categories. The three primary categories were administrative (n=203, 43.1%), provider support (n=159, 33.8%), and patient care (n=109, 23.1%). Within administrative apps, we split the apps into four secondary categories: front office (n=77, 37.9%), financial (n=53, 26.1%), office administration (n=49, 24.1%), and office device integration (n=17, 8.4%). Within the provider support primary classification, we split the apps into eight secondary categories: documentation (n=34, 21.3%), records management (n=27, 17.0%), care coordination (n=23, 14.4%), population health (n=18, 11.3%), EHR efficiency (n=16, 10.1%), ordering and prescribing (n=15, 9.4%), medical device integration (n=13, 8.2%), and specialty EHR (n=12, 7.5%). Within the patient care primary classification, we split the apps into three secondary categories: patient engagement (n=50, 45.9%), clinical decision support (n=40, 36.7%), and remote care (n=18, 16.5%). Total app counts varied substantially across EHR vendors. Overall, the distribution of apps across primary categories were relatively similar, with a few exceptions.

Conclusions: We characterized and organized a diverse and rich set of third-party EHR apps. This work provides an important reference for developers, researchers, and EHR customers to more easily search, review, and compare apps in EHR app marketplaces.

(JMIR Med Inform 2020;8(5):e16980) doi:10.2196/16980

KEYWORDS

electronic health records; medical informatics; software; interoperability; apps; app marketplace

Introduction

The electronic health record (EHR) stores patient health information, automates clinical workflows, and supports other care-related functions such as clinical decision support [1]. Clinical and governmental drivers have facilitated widespread adoption of EHRs in health care worldwide [2]. Health care providers rely on EHRs to perform essential functions such as documenting patient encounters, providing clinical decision support, and engaging patients in their own care [3,4]. However, EHR implementation hurdles, usability flaws, and poor interoperability, among other issues, keep EHRs from delivering their full potential benefit to health care organizations [5]. The

variation in EHR implementation across health organizations contributes to these problems as each organization may rely on different methods to integrate additional value into their EHR systems. Some organizations may leverage custom integration with third-party applications whereas others may resort to in-house development or other integration strategies to support their needs [6,7]. In any case these integrations tend to be time-consuming, expensive, and limited for use only within their respective organizations [8]. Ideally, a successful information technology (IT) application that integrates with an EHR at one organization would be available for the same integration with an EHR at another organization, regardless of EHR vendor [9].

The EHR app model, inspired by smartphone app marketplaces, has been proposed to increase flexibility and availability of EHR integrations while also fostering innovation in health IT [9-11]. This approach is made possible by increased EHR interoperability and standardized access to EHR data through application programming interfaces (APIs) and standards such as FHIR (Fast Healthcare Interoperability Resources) [12]. One early implementation, the SMART (Substitutable Medical Applications, Reusable Technologies) App Gallery [13] is an example of an EHR app platform that heavily leverages the FHIR standard to enable a plug-and-play style of integration with participating EHRs [14]. SMART app development depends heavily on two major concepts—apps must be both substitutable and reusable. A substitutable app accesses EHR data and can be easily added, replaced, or deleted within an EHR. This allows health care organizations to choose the app that best fits their needs [15]. A reusable app is developed once but can be installed by many clients potentially across multiple EHRs [15]. The FHIR standard serves as the common data specification that both EHR vendor APIs and SMART APIs adhere to in order to interoperate. An EHR app platform built on these concepts increases access to health IT solutions for health care organizations and allows third-party app developers to compete in a market driven by the value and price of their app [9]. Motivated by these benefits [16], major EHRs have started to create their own app marketplaces to encourage development of apps on their own platforms [15]. There are now hundreds of apps available on EHR app marketplaces.

Given the widespread utilization of EHRs and the emergence of EHR app marketplaces, it has become necessary to conduct a systematic search and analysis of apps in EHR app marketplaces in an effort to help organize, categorize, and characterize available EHR apps. This study will help health professionals, researchers, and developers understand the currently available technologies; make it easier to review, search, and compare available EHR apps; provide a common vocabulary to facilitate communication; identify where gaps and opportunities exist for research and development; and justify investment into the research and development of new EHR apps.

Methods

App Extraction

We identified and reviewed all known apps in public-facing marketplaces of the top 10 EHR vendors in the United States, which include (in order of market share) Epic Systems Corporation; Allscripts; eClinicalWorks, LLC; NextGen Healthcare; GE Healthcare; athenahealth, Inc; Cerner Corporation; Greenway Health LLC; Practice Fusion; and eMDs [17]. GE Healthcare did not have a public facing app marketplace at the time of this writing and Practice Fusion was recently acquired by Allscripts [18]. The leading vendors in the US market were chosen for analysis because they had publicly available app marketplaces (with the exception of GE Healthcare), and they represented a cluster of vendors serving the majority of a common set of customers.

We used custom web scrapers and public ReST (Representational State Transfer) endpoints from these EHR marketplaces to gather EHR app data such as name, description, links, website, EHR-defined app categories, ratings, reviews, EHR versions, and other available information. This information was recorded in our EHR app database wherein we consolidated duplicate EHR apps that were listed in multiple EHR marketplaces. The last data extraction occurred in February 2019. Clear indication of FHIR compatibility was not consistently present in the extracted data within or across EHR marketplaces. Marketplace offerings that offered professional services without clear evidence of EHR integration (eg, website builders or marketing services), teams of professionals granted access to EHR interfaces (eg, offsite medical coders), and medical devices without EHR integration (eg, stand-alone weight scale) were not considered apps and were excluded.

App Review

Two informaticists (authors JR and BW) used grounded theory principles to create categories that emerged from the EHR app information [19]. EHR app classifications were created inductively by each reviewer independently based upon available information about each EHR app. Importantly, not all data fields were available across all EHRs. For example, not all marketplaces included information indicating whether the app was open source or a commercial offering. Even though some marketplaces included this information, apps were generally classified based on the information available that was common across apps from all marketplaces. When the EHR app information was either inadequate or missing, making it difficult to accurately classify the app, we referenced the app developer's website. As common EHR app features, functions, or purposes emerged, we created categories to group similar apps. Where similarity between categories existed, they were related to form larger, more inclusive categories. Conversely, if sufficient EHR app divergence existed within a category, we subclassified the apps into more unified categories [20]. We drew category names from app descriptions, EHR-designated classifications, and common industry concepts. A minimum of three apps were required to form a category.

Following the review and initial classification by each reviewer independently, a joint review process commenced between the two reviewers. The reviewers worked together to come to a consensus on category names, the organization of the taxonomic class hierarchy, and the correct classification of each EHR app. To facilitate the joint review process, a database with all EHR app information as well as notes from each reviewer and their

initial classifications were utilized. Each EHR app was reviewed and discussed together. As consensus categories emerged, formal definitions for each category were created and refined. A final classification was used to denote the consensus classification. Consensus was reached when both reviewers agreed with the classification. In cases where the reviewers failed to reach consensus, a third-party arbitrator was available. Through multiple rounds of discussion and debate, a consensus EHR app taxonomy emerged.

Results

App Extraction

Of the eight EHRs with public-facing marketplaces, we identified a total of 749 offerings. The total number of offerings for marketplaces ranged from 21 (eMD) to 227 (Athenahealth). In total, 153 offerings were listed on at least two EHR marketplaces, resulting in a total of 596 unique offerings; 125 were excluded from consideration for not meeting our inclusion criteria, which resulted in 471 unique apps being incorporated into our taxonomy. We categorized the EHR app into a taxonomy consisting of 3 primary categories, 15 secondary categories, and 55 tertiary categories. The three primary categories were administrative (n=203 apps, 43.1%), provider support (n=159, 33.8%), and patient care (n=109, 23.1%).

For each EHR, the distribution of apps across the primary categories followed a similar trend. In general, administrative apps make up the greatest portion of EHR apps, followed by provider support apps and then primary care apps. Interestingly, Cerner's marketplace has a higher ratio of patient care apps, followed by provider support and then administrative. There was also a large variability in the number of listings excluded for not meeting criteria for being an app between EHR marketplaces, with eClinicalWorks accounting for 65 of 125 (52%) excluded offerings (Table 1). Each primary category is described in further detail below.

Table 1. Distribution of app marketplace offerings across primary categories by electronic health record (EHR) vendors.

EHR vendor	Primary category			
	Administrative, n (%)	Provider support, n (%)	Patient care, n (%)	Not an app, n (%)
Athenahealth (n=222)	90 (40.5)	63 (28.4)	45 (20.3)	24 (10.8)
eClinicalWorks (n=116)	23 (19.8)	14 (12.1)	14 (12.1)	65 (56.0)
Epic (n=113)	46 (40.7)	33 (29.2)	30 (26.6)	4 (3.5)
Allscripts (n=110)	33 (30.0)	47 (42.7)	28 (25.5)	2 (1.8)
Greenway (n=87)	48 (55.2)	19 (21.8)	4 (4.6)	16 (18.4)
Nextgen (n=37)	10 (27.0)	14 (37.8)	4 (10.8)	9 (24.3)
Cerner (n=28)	1 (3.6)	10 (35.7)	17 (60.7)	0 (0.0)
eMD (n=21)	9 (42.9)	3 (14.3)	0 (0.0)	9 (42.9)

App Review

Administrative

The 203 administrative apps facilitate the administrative functions of a hospital or clinic. Within this classification, we split the apps into four secondary categories: front office (n=77, 37.9%), financial (n=53, 26.1%), office administration (n=49, 24.1%), and office device integration (n=17, 8.4%). The administrative app categories, descriptions, and counts are listed in Table 2.

Table 2. Category definitions and counts for administrative apps (n=203).

Categories	Category descriptions	Count, n (%)
Administrative	**Facilitates the conduct of administrative functions of a hospital or clinic setting**	**203 (100.0)**
Front office	**Helps support front office staff interaction with patients**	**77 (37.9)**
Scheduling	Helps schedule and manage patient appointments	32 (15.7)
Patient check-in	Helps manage the patient check-in process	12 (5.9)
Patient communication	Facilitates communication with patient for administrative purposes	12 (5.9)
Document management	Helps capture and manage documents	10 (4.9)
Answering service	Captures information related to after-hours patient calls	4 (2.0)
Phone triage	Facilitates triage according to industry standard protocols	3 (1.5)
Financial	**Helps manage the financial needs of the clinic**	**53 (26.1)**
Patient billing	Captures and processes payment information from patients	20 (9.9)
Insurance	Facilitates claims and authorization	13 (6.4)
Collections	Manages patient collections	8 (3.9)
Medical coding	Improves accuracy and efficiency of medical coding	7 (3.4)
Patient pay estimation	Estimates cost of care	3 (1.5)
Office administration	**Supports the administrative needs of the clinic**	**49 (24.1)**
Analytics and reporting	Helps track, analyze, and report on clinical operations	17 (8.4)
Patient experience	Measures the clinical experience of the patient	13 (6.4)
Inventory management	Tracks inventory of medical products	7 (3.4)
IT[a] systems management	Supports the IT system needs of a health care organization	7 (3.4)
Compliance	Helps maintain, track, and/or report compliance	4 (2.0)
Office device integration	**Device used by office staff for administrative purposes**	**17 (8.4)**
Scanner integration	Integrates scanners with the EHR[b]	7 (3.4)
Printer integration	Integrates printers with the EHR	3 (1.5)
Signature pad integration	Integrates a signature pad with the EHR	3 (1.5)

[a]IT: information technology.

[b]EHR: electronic health record.

Provider Support

We identified 159 provider support apps, which we defined as apps that primarily support the functions of care providers in their delivery of health care to patients. Within the provider support primary classification, we split the apps into eight secondary categories: documentation (n=34, 21.3%), records management (n=27, 17.0%), care coordination (n=23, 14.4%), population health (n=18, 11.3%), EHR efficiency (n=16, 10.1%), ordering and prescribing (n=15, 9.4%), medical device integration (n=13, 8.2%) and specialty EHR (n=12, 7.5%). The provider support app categories, descriptions, and counts are in Table 3.

Table 3. Category definitions and counts for provider support apps (n=159).

Categories	Category descriptions	Count, n (%)
Provider support	**Supports the functions of care providers in their delivery of health care**	**159 (100.0)**
Documentation	**Facilitates the collection and management of patient information**	**34 (21.3)**
Dictation and transcription	Transcribes dictated clinical narratives into clinical notes	12 (7.5)
Structured documentation	Facilitates efficient and accurate documentation	7 (4.4)
Image capture	Captures images for documentation (usually a mobile device)	6 (3.8)
Natural language processing	Uses natural language processing to process unstructured data	6 (3.8)
Records management	**Supports access to or management of records**	**27 (17.0)**
Image management	Allows access to or management of images, including RIS/PACS[a] systems	11 (6.9)
Legacy/migration	Provides access to legacy medical records or facilitates the conversion of paper records to electronic records	8 (5.0)
Access	Consolidates patient records in one view or allows access via mobile device	5 (3.1)
Backup	Saves data in an alternate form that can be accessed in the event of an outage	3 (1.9)
Care coordination	**Helps care team members coordinate their care for a patient**	**23 (14.4)**
Clinic scheduling	Manages the scheduling and workflow of providers in a clinic	7 (4.4)
Service directory	Provides a list of services or providers to refer or access	7 (4.4)
Provider communication	Facilitates the communication between care team members about a patient	6 (3.8)
Population health	**Helps manage that health of a population or group of patients**	**18 (11.3)**
Chronic care management	Helps providers manage chronic conditions in patients	10 (6.3)
Annual wellness visit	Facilitates annual wellness visit scheduling and reporting	4 (2.5)
Population risk assessment	Helps identify and manage at-risk patients	4 (2.5)
EHR[b] efficiency	**Makes the EHR easier or more efficient for the provider to use**	**16 (10.1)**
Information display	Consolidates patient record into easily consumed dashboards, reports, and infographics	9 (5.7)
Ordering and prescribing	**Facilitates the ordering or prescribing of a device, substance, or service**	**15 (9.4)**
Prescription drug monitoring program	Provide access to state Prescription Drug Monitoring Program databases	5 (3.1)
Pharmacy	Manages electronic prescription renewal and ordering	4 (2.5)
Medical equipment	Manages electronic ordering of DME[c]	3 (1.9)
Image ordering	Facilitates image ordering or helps manage image ordering workflow	3 (1.9)
Medical device integration	**Device used by health care provider for clinical purposes**	**13 (8.2)**
Cardiac devices	Collects data from cardiac devices	5 (3.1)
Digital scales	Collects data from digital scales	3 (1.9)
Specialty EHR	**Extends the functions of an EHR to support a specific clinical domain or specialty**	**12 (7.5)**
Obstetrics	Extends the EHR to provide functionality for prenatal and perinatal data management	3 (1.9)
Anesthesia	Extends the EHR to provide functionality for Anesthesia data management	3 (1.9)

[a]RIS/PACS: Radiology Information System/Picture Archiving and Communication System.

[b]EHR: electronic health record.

[c]DME: durable medical equipment.

Patient Care

The 109 patient care apps we identified facilitate the provision of clinical care between a health care provider and a patient. Within the patient care primary classification, we split the apps into three secondary categories: patient engagement (n=50, 45.9%), clinical decision support (n=40, 36.7%), and remote care (n=18, 16.5%). The patient care app categories, descriptions, and counts are listed in Table 4.

Table 4. Category definitions and counts for patient care apps (n=109).

Categories	Category descriptions	Count, n (%)
Patient care	**Facilitates the provision of clinical care between a health care provider and a patient**	**109 (100.0)**
Patient engagement	**Engages patients in their own health care**	**50 (45.9)**
Patient assessment	Collects patient-reported information for a clinical purposes	15 (13.8)
Care plan management	Helps patient follow a provider's rehab instructions, medication schedule, and/or care plan regimen	10 (9.2)
Patient education	Provides education and instruction resources to patients specific to their care	9 (8.3)
Health record access	Allows patients to access, download, and share their medical records, or allows providers to fulfill medical record requests	6 (5.5)
Patient wearables	Records information from patient wearables (passive involvement)	6 (5.5)
Clinical decision support	**Provides or delivers decision support to providers based on patient data**	**40 (36.7)**
Ordering CDS[a]	Supports the appropriateness of medication, imaging, and lab test orders	10 (9.2)
Medication CDS	Provides decision support for medication dosing and monitoring	9 (8.3)
Patient risk assessment	Assess a patient's health risk	5 (4.6)
Knowledge management	Provides access to and manages medical knowledge for providers	5 (4.6)
Patient monitoring	Monitors health of patient and alerts provider of notable changes	4 (3.7)
Remote care	**Supports the provision of care to patient remotely**	**18 (16.5)**
Telehealth platform	Gives a provider technical capabilities to meet with a patient remotely	12 (11.0)
Remote consult	Provides access to care providers or specialists who are remote	5 (4.6)

[a]CDS: clinical decision support.

Discussion

Principal Results

We conducted a systematic search and analysis of apps in EHR app marketplaces to help organize, categorize, and characterize available EHR apps. This study brings value to the health IT industry because it creates a common vocabulary that can be used to communicate about EHR apps, helps health care organizations identify EHR app solutions, and justifies investment into the research and development of new EHR apps. With this study, we can identify common patterns of EHR integration approaches and create a template to streamline future EHR app development and integration. This helps researchers identify gaps in integration capabilities, standards, or functionalities that need to be addressed by governing bodies, standards organizations, EHR vendors, or EHR app developers.

Our EHR app review organized and characterized 471 unique EHR apps into 3 primary categories, 15 secondary categories, and 55 tertiary categories. Several categories were larger or more well-defined than others. Administrative apps represented the largest share of EHR apps with 203 apps. Provider support and patient care apps were the other primary categories with 159 and 109 EHR apps, respectively. EHR marketplaces tended to reflect this overall trend with the majority of apps falling under the administrative category, followed by provider support and patient care. Cerner followed a distinctly different trend with patient care representing the majority of their apps and only a single administrative app. This may in part be attributed to how Cerner validates apps that are submitted by third-party developers. While all app galleries reviewed here have a submission process, and several list disclaimers that not all submissions may be listed upon submission, Cerner has an additional validation step. Apps that don't meet a certain standard set by Cerner may be rejected or asked to resubmit after outstanding issues are resolved. This adds extra rigor in the Cerner app submission process that may account for the lower total app count in their gallery as well as the different ratio of primary categories observed. The fact that zero offerings in Cerner's gallery were considered "Not an app" and excluded from consideration in our search and analysis may be attributed to their unique validation approach and, generally speaking, Cerner's apps required less attention when assigning apps to categories. However, Cerner also had fewer total offerings listed in their gallery (n=28) than all other vendors except for eMD (n=21). App quality is an important issue and while the right amount of validation is difficult to quantify [8], it is important to note that if validation is too strict, it could suppress innovation and defeat a key purpose of the app model, which allows competition among developers based on app value and price [9]. This allows clients to validate app offerings and reward innovation and the value the apps provide [15].

Interestingly, provider support apps had the greatest variability and ambiguity among the three primary categories of apps. Provider support had more secondary categories than the other two primary categories combined, accounting for 8 of the 15 (53%) total secondary categories. Many apps offered multiple functionalities or had feature sets that made it difficult to assign secondary and tertiary categories. The value these apps provided and how they were intended to be used by the provider was often unclear. Provider support accounted for 63% (14/22) of

apps that were not specific enough to place in a secondary category. Additionally, provider support accounted for 46% (25/54) of apps that were not specific enough to place in a tertiary category. The secondary categories under provider support with the most apps without tertiary categories were specialty EHR (6/12, 50%), EHR efficiency (7/16, 44%), and medical device integration (7/13, 39%). Among all other secondary categories, office device integration had the highest percentage of apps that did not fall into a tertiary category (4/17, 24%), followed by clinical decision support (7/40, 18%). This suggests that the provider support category has the greatest need for further refinement and innovation out of the three primary categories.

The app model is a remedy to the one-size-fits-all strategy that is failing to meet the needs of patients, providers, and administrators in a rapidly evolving landscape [8]. For the app model to be effective, the apps listed need to solve a clearly defined problem instead of offering diverse sets of features and functionalities that begin to approximate a one-size-fits-all-solution. In other words, it needs to be obvious what category the app belongs in. The taxonomy of apps we have curated will help health IT companies and app developers match app development to a well-defined purpose and assist health professionals in identifying gaps in the current set of app categories. As app functionality and feature sets become more cohesive and achieve alignment with specific EHR app categories, the value and impact the app model will have on health care will increase as patients, providers, and administrators can more easily search, install, and ultimately be the market force that will drive innovation and value of EHR apps.

Limitations

We acknowledge several shortcomings of the current review. First, our review is limited only to apps currently available on the top 10 EHR app marketplaces by market share in the United States. We acknowledge there are other EHR vendors worldwide developing app marketplaces as well that were not reviewed here. During our review and research of EHR apps, we came across several other apps that claimed EHR integration that would have been included if they had been listed in an EHR marketplace. It would be unfeasible to know all apps that integrate with EHRs; nevertheless, we hope that as EHR marketplaces mature, these apps will become listed in the EHR marketplaces and organized in our review. Second, several apps had characteristics or features that could justify their classification under more than one category. In these cases, we endeavored to classify apps to the lowest level in the taxonomy as possible while still accurately reflecting the apps' primary purpose, which sometimes resulted in higher level classifications. In a few instances, when the information was insufficient to determine whether an offering was an app, we erred on the side of inclusion. As a result, a few apps in our taxonomy may have been inappropriately included. As further information becomes available, their inclusion and classification will be re-evaluated. Third, the review was conducted by two informaticists. We recognize that shortcomings, inaccuracies, and/or bias may exist in the interpretation and characterization of the apps. Independent input from a panel of expert stakeholders would increase robustness and validity of the EHR app review. Finally, the EHR app review represents a single point in time (February 2019). However, as new EHR apps get added to marketplaces and new app information becomes available, the results will become outdated. We anticipate conducting this review again in a few years to understand how EHR app marketplaces have evolved over time.

Comparison With Prior Work

The SMART app model was proposed in 2009 by Mandl et al [9]. Since then, EHR vendors have followed suit by building their own app marketplaces. We identified hundreds of apps in these marketplaces that allow integration with their respective EHR vendor. Current EHR marketplaces do not fully reflect the original proposal made by Mandl et al [9], which called for total substitutability of apps across any EHR by conforming to a single standard. Without conforming to a single standard as proposed by Mandl et al [9], each app would need to integrate with each EHR marketplace individually, as is the case today. For instance, we observed that 153 apps integrate with two or more EHRs. This does not quite meet the proposal made by Mandl et al [9] where an app lives on a single platform and can be integrated with any health system regardless of EHR vendor.

Conclusions

We characterized and organized a diverse and rich set of third-party EHR apps. This work provides an important reference for developers, researchers, and EHR customers to more easily search, review, and compare apps in EHR app marketplaces. While future research and validation among independent informaticists and stakeholders will increase the validity and value of this review, this work provides a strong foundation upon which future EHR app research will be established.

Conflicts of Interest

None declared.

References

1. Romano MJ, Stafford RS. Electronic health records and clinical decision support systems: impact on national ambulatory care quality. Arch Intern Med 2011 May 23;171(10):897-903 [FREE Full text] [doi: 10.1001/archinternmed.2010.527] [Medline: 21263077]
2. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. Health Aff (Millwood) 2017 Aug 01;36(8):1416-1422. [doi: 10.1377/hlthaff.2016.1651] [Medline: 28784734]

3. Samal L, Linder JA, Lipsitz SR, Hicks LS. Electronic health records, clinical decision support, and blood pressure control. Am J Manag Care 2011 Sep;17(9):626-632 [FREE Full text] [Medline: 21902448]
4. Burke HB, Sessums LL, Hoang A, Becher DA, Fontelo P, Liu F, et al. Electronic health records improve clinical note quality. J Am Med Inform Assoc 2015 Jan;22(1):199-205 [FREE Full text] [doi: 10.1136/amiajnl-2014-002726] [Medline: 25342178]
5. Kalorama Information. EMR Market 2017: Electronic Medical Records in an Era of Disruption URL: https://kaloramainformation.com/product/emr-market-2017-electronic-medical-records-in-an-era-of-disruption/ [accessed 2019-03-13]
6. Health IT Outcomes. EHR Failure: What's A Practice To Do? URL: https://www.healthitoutcomes.com/doc/ehr-failure-what-s-a-practice-to-do-0001 [accessed 2019-03-12]
7. Healthcare IT News. 2012. 12 integration capabilities EHRs will need to have URL: https://www.healthcareitnews.com/news/12-integration-capabilities-ehrs-will-need-have [accessed 2019-03-12]
8. Mandl K, Mandel J, Pfiffner P. Chapter 16 - An Apps-Based Information Economy in Healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, editors. Key Advances in Clinical Informatics. Cambridge, Massachusetts: Academic Press; 2017:227-236.
9. Mandl KD, Kohane IS. No small change for the health information economy. N Engl J Med 2009 Mar 26;360(13):1278-1281. [doi: 10.1056/NEJMp0900411] [Medline: 19321867]
10. Mandl KD, Kohane IS. Escaping the EHR trap--the future of health IT. N Engl J Med 2012 Jun 14;366(24):2240-2242. [doi: 10.1056/NEJMp1203102] [Medline: 22693995]
11. Mandl K, Kohane I, Christensen C, Chueh H, Frisse M, Kibbe D. smarthealthit.ork. 2009. Ten Principles for Fostering Development of an "iPhone-like" Platform for Healthcare Information Technology URL: http://smarthealthit.org/wp-content/uploads/CHIP-HIT-Platform.pdf [accessed 2020-05-12]
12. Sittig DF, Wright A. What makes an EHR "open" or interoperable? J Am Med Inform Assoc 2015 Sep;22(5):1099-1101 [FREE Full text] [doi: 10.1093/jamia/ocv060] [Medline: 26078411]
13. smarthealthit.org. SMART App Gallery URL: https://gallery.smarthealthit.org/ [accessed 2019-03-12]
14. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc 2016 Feb 17 [FREE Full text] [doi: 10.1093/jamia/ocv189] [Medline: 26911829]
15. Mandl KD, Mandel JC, Kohane IS. Driving Innovation in Health Systems through an Apps-Based Information Economy. Cell Syst 2015 Jul;1(1):8-13 [FREE Full text] [doi: 10.1016/j.cels.2015.05.001] [Medline: 26339683]
16. healthit.gov. Connecting Health and Care for the Nation: A Shared Nationwide Interoperability Roadmap URL: https://www.healthit.gov/sites/default/files/hie-interoperability/nationwide-interoperability-roadmap-final-version-1.0.pdf [accessed 2020-05-12]
17. Health IT Dashboard. Health Care Professional Health IT Developers URL: https://dashboard.healthit.gov/quickstats/pages/FIG-Vendors-of-EHRs-to-Participating-Professionals.php [accessed 2019-03-12]
18. healthcareitnews.com. 2018. Allscripts buys Practice Fusion for $100 million URL: https://www.healthcareitnews.com/news/allscripts-buys-practice-fusion-100-million [accessed 2019-06-24]
19. Charmaz K. Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. New York City, NY: SAGE Publications; 2006.
20. Rasch R. The nature of taxonomy. Image J Nurs Sch 1987;19(3):147-149. [doi: 10.1111/j.1547-5069.1987.tb00613.x] [Medline: 3666771]

Abbreviations

API: application programming interface
CDS: clinical decision support
DME: durable medical equipment
EHR: electronic medical record
FHIR: Fast Healthcare Interoperability Resources
IT: information technology
ReST: Representational State Transfer
RIS/PACS: Radiology Information System/Picture Archiving and Communication System
SMART: Substitutable Medical Applications, Reusable Technologies

Edited by C Lovis; submitted 08.11.19; peer-reviewed by K Mandl, C Fincham, M Kolotylo-Kulkarni, N Sakib; comments to author 09.02.20; revised version received 20.03.20; accepted 12.04.20; published 29.05.20.

Please cite as:
Ritchie J, Welch B
Categorization of Third-Party Apps in Electronic Health Record App Marketplaces: Systematic Search and Analysis
JMIR Med Inform 2020;8(5):e16980
URL: http://medinform.jmir.org/2020/5/e16980/
doi:10.2196/16980
PMID:32469324

Original Paper

Effect of Online Health Information Seeking on Anxiety in Hospitalized Pregnant Women: Cohort Study

Fabiana Coglianese[1], MSc; Giulia Beltrame Vriz[2], MSc; Nicola Soriani[3], PhD; Gianluca Niccolò Piras[3], MD; Rosanna Irene Comoretto[3], PhD; Laura Clemente[1], BSc; Jessica Fasan[1], MSc; Lucia Cristiano[1], BSc; Valentina Schiavinato[4], PhD; Valter Adamo[1], MD; Diego Marchesoni[5], MD; Dario Gregori[3], MA, PhD

[1]Unit of Obstetrics and Gynecology, Maternal-Infant Department, Santa Maria degli Angeli Hospital, Pordenone, Italy

[2]Department of Obstetrics, Burlo Garofolo Pediatric Institute, Trieste, Italy

[3]Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padua, Padua, Italy

[4]Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Padua, Italy

[5]Unit of Obstetrics and Gynecology, Maternal-Infant Department, Santa Maria della Misericordia Hospital, Udine, Italy

Corresponding Author:
Dario Gregori, MA, PhD
Department of Cardiac, Thoracic, Vascular Sciences and Public Health
University of Padua
Via Loredan 18
Padua, 35121
Italy
Phone: 39 0498275384
Email: dario.gregori@unipd.it

Abstract

Background: There are approximately 1,000,000 pregnant women at high risk for obstetric complications per year, more than half of whom require hospitalization.

Objective: The aim of this study was to determine the relation between online health information seeking and anxiety levels in a sample of hospitalized woman with pregnancy-related complications.

Methods: A sample of 105 pregnant women hospitalized in northern Italy, all with an obstetric complication diagnosis, completed different questionnaires: Use of Internet Health-information (UIH) questionnaire about use of the internet, EuroQOL 5 dimensions (EQ-5D) questionnaire on quality of life, State-Trait Anxiety Inventory (STAI) questionnaire measuring general anxiety levels, and a questionnaire about critical events occurring during hospitalization.

Results: Overall, 98/105 (93.3%) of the women used the internet at home to obtain nonspecific information about health in general and 95/105 (90.5%) of the women used the internet to specifically search for information related to their obstetric disease. Online health information-seeking behavior substantially decreased the self-reported anxiety levels (P=.008).

Conclusions: Web browsing for health information was associated with anxiety reduction, suggesting that the internet can be a useful instrument in supporting professional intervention to control and possibly reduce discomfort and anxiety for women during complicated pregnancies.

(JMIR Med Inform 2020;8(5):e16793) doi:10.2196/16793

KEYWORDS

anxiety; pregnant women; web health information; internet use

Introduction

Approximately 1,000,000 pregnant women are at high risk for obstetric complications globally per year, about 700,000 of whom will require hospitalization. Preterm labor, placenta previa, pregnancy-induced hypertension, and gestational diabetes are some of the most common conditions during pregnancy that require medical attention [1,2].

Moreover, mental disorders can affect the pregnancy course, especially in high-risk pregnancies, which can exacerbate depression and anxiety, and hospitalization can further increase stress levels [3,4]. According to the World Health Organization,

mental health disorders are the leading cause of disease burden in women aged between 15 and 44 years [5], corresponding to the main fertile window. In this regard, depression or anxiety during pregnancy has been associated with poor maternal health behaviors (eg, tobacco use) and with adverse birth outcomes (eg, preterm labor). Moreover, anxiety or depression during pregnancy may also adversely affect the development of the infant/child [6-8].

Antenatal depression and anxiety occur in approximately 13% and 21.7% of women, respectively [9]. The former affects 19% of women hospitalized for obstetric risk [3], whereas about 1 out of 3 pregnant women suffers from anxiety. In particular, the prevalence of depression and anxiety is higher in the first and third trimester (36.3% and 35.8%, respectively) and is slightly lower during the second trimester (32.3%) [3,10].

Despite the relevance for pregnancy outcome, mental health of the mother, and development of the child, few studies have directly explored depression, anxiety, quality of life, and possibilities of mental health treatment in women hospitalized for high-risk pregnancies. The coexistence of anxiety and depression in this vulnerable group reaches up to 40%, which is 3 times greater than the rate reported in community-based samples of pregnant women [10].

Online health information-seeking behavior has become increasingly popular among pregnant women owing to the several uncertainties that can arise during pregnancy [11-14]. Moreover, health care professionals often provide pregnant women with informational support, especially underlining where and how to obtain the resources they need [15,16]. Although many studies have investigated the psychological and environmental factors predisposing subjects to online health information-seeking behaviors, only few have examined the effect of this behavior and its relationship with anxiety [16].

Therefore, the aim of this study was to determine the relationship between online health information-seeking behaviors and anxiety levels among a sample of women hospitalized for a pregnancy-related issue. Moreover, we aimed to understand how anxiety levels change during hospitalization by comparing anxiety levels and access to online health information between women having access to the internet during hospitalization and those who did not.

Methods

Study Design and Setting

We performed a two-center cohort study at the Departments of Obstetrics of Santa Maria della Misericordia Hospital in Udine and Santa Maria degli Angeli Hospital in Pordenone, Italy between August 2015 and March 2016. Women enrolled in the study were >18 years old and 1-40 weeks pregnant who were hospitalized for obstetrics-related complications, including gestational diabetes, preeclampsia, pregnancy-induced hypertension, renal colic, and severe hyperemesis. Women without pathological pregnancy, with cognitive or major psychiatric diseases, and nonnative speakers of Italian were excluded from the study. To minimize bias, both nursing and midwifery staff were instructed not to interfere with patients' spontaneous usage of internet resources.

Data collection was based on electronic case report forms maintained on the Research Electronic Data Capture [17] system of the Service for Clinical Trials and Biometrics of the Unit of Biostatistics, Epidemiology and Public Health (Department of Cardiac, Thoracic, Vascular Sciences and Public Health, University of Padova, Italy). The study received authorization by the Region Ethical Committee (CERU; protocol 17002, Opinion no. 37/2015, 7/7/2015). Informed consent was obtained from all individual participants included in the study.

Data Collection

The pregnant women enrolled in the study were asked to fill out various questionnaires during hospitalization: Use of Internet Health-information (UIH) questionnaire about use of the internet, EuroQOL 5 dimensions (EQ-5D) questionnaire on quality of life, State-Trait Anxiety Inventory (STAI) questionnaire for measuring two distinct anxiety concepts, and a questionnaire about critical events occurring during hospitalization. Additional demographic and clinical information, including gender, age, education level, obstetric history (childbirth, miscarriage, ectopic pregnancy, and type of pregnancy problem), and use of alcohol and tobacco, were collected by the study researchers based on medical histories. Critical events such as medical complications, hospital dissatisfaction, and family problems occurring during the hospitalization period were also recorded.

Internet Health Information Questionnaire

The UIH questionnaire on online information-seeking behaviors [18-20] was adapted for this study population. This questionnaire is divided into 3 parts. The first part investigates internet usage at home, patients' attitude with respect to searching health information, the type of information most frequently searched for, the general frequency of web use, and the tendency to share this information with health care providers, usually a midwife. This part of the questionnaire was administered only at the beginning of hospitalization. The second part, which was administered every day until discharge, investigated internet usage during the hospital stay, the tool used (eg, smartphone, tablet, notebook), and the time spent searching for information about their health condition. The third part was composed of a visual analog scale (UIH-VAS) regarding the amount of time spent on the internet in the last unit of time (usually the day) in searching for information about a specific disease or regarding general health-related information.

EuroQOL 5 Dimensions Health-Related Quality of Life Questionnaire

The EQ-5D questionnaire [21] was adopted to measure health-related quality of life, which consists of a questionnaire and a visual analogue scale (EQ-VAS). The EQ-VAS records subjects' perceptions of their own current health status and can be used to monitor changes over time. The questionnaire is a self-reported description of subjects' current health in five dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Subjects are asked to grade their own current level of function in each dimension choosing between

three degrees (severe, moderate, or none). Combining the different information, 245 distinct health states can be described.

State-Trait Anxiety Inventory

The STAI [22], which was adapted for the Italian population [23], is a questionnaire frequently used in pregnant women affected by obstetric diseases to evaluate nonpathological anxiety levels. The STAI is composed of two self-reported scales for measuring two distinct anxiety concepts: state anxiety and trait anxiety. Both scales contain 20 statements that ask the respondents to describe how they feel at a given time (state anxiety) and how they generally feel (trait anxiety). In this way, state anxiety is conceptualized as a transitory emotional state, whereas trait anxiety refers to relatively stable individual differences in their propensity for anxiety. The state anxiety questionnaire was administered every day until discharge, whereas the trait anxiety questionnaire was filled out only once at the beginning of hospitalization.

Sample Size Calculation

This research was powered to detect potential differences on the average STAI score (range 20-80) of 6 points between women with internet access (with a minimum of 10 minutes/day of web browsing, excluding emails) and those without internet access. Based on previous estimates using the same instrument [24], assuming an SD of 8 points in the differences of STAI scores and assuming a ratio of 0.42 between the rate of women with and without internet access (for α=.03 and 1 – β=.85), a total of 105 pregnant women were planned to be recruited (using a two-sample *t* test with unknown variance).

Statistical Analysis

Descriptive data are presented as the median (IQR) for continuous variables and as absolute numbers (percentages) for categorical variables as appropriate. Unadjusted differences were tested using Wilcoxon or Chi square tests without continuity corrections as appropriate depending on the variable analyzed.

Effects of relevant confounders on STAI scores and internet usage were considered by estimates in a multivariate longitudinal linear model. The marginal effects of relevant covariates were estimated using the Huber-White sandwich estimator and an autoregressive correlation structure [25]. Variables were selected from a pool of significant variables based on univariate analyses according to an Akaike information criterion value at least 0.25 [26] in a forward fashion with a significance threshold of P=.10. Age and quality of life, measured by the EQ-VAS, were forced to stay in the model regardless of their significance. Nonlinear effects of covariates were estimated using restricted cubic splines and their significance was estimated using a log-likelihood ratio test. A specific term for the interaction between UIH and time was added to the final model to evaluate its statistical significance and was eventually removed if the corresponding P value fell below .05. Goodness of fit was evaluated using the R^2 value on a set of bootstrapped (B=10,000) resamples. The analysis was performed using the RMS libraries [27] and R software packages [28].

Results

A total of 105 hospitalized pregnant women were recruited for the study. The main characteristics of the study sample are provided in Table 1, stratified by internet usage group. Overall, the preferred tool for internet use was a personal computer.

With respect to internet usage, 98 out of 105 women (93.3%) reported using the internet at home not only for emails but also to seek health-related information. The majority of women were looking for health or medical information for themselves (95/105, 90.5%) or for someone else (70/105, 66.7%); to request personal health information such as test results or medical appointments (72/105, 68.6%); to communicate with physicians (55/105, 52.4%); or to consult informational websites about weight, diet, or physical activity during pregnancy (36/105, 34.3%). Moreover, 95 of the 105 subjects (90.5%) used internet specifically to obtain information on their obstetric disease: 85 of 95 women (89%) found the information useful, whereas only 44 of 93 participants (47%) shared the information they found with their health care providers. Only 7 of the total 105 women (6.7%) had not been using the internet at home.

Internet use was virtually absent after the first two days of hospitalization (Figure 1). Therefore, behaviors of internet use and the outcome variables are presented in Table 2 only for the first two days of hospitalization and at discharge. Statistically significant differences within the two subgroups (with and without internet use) were observed on the UIH-VAS scale.

Table 1. Sample characteristics stratified by internet usage for seeking health-related information at home.

	N	Did not use the internet before hospitalization (n=7)	Used the internet before hospitalization (n=98)	All subjects	*P* value
Age (years), median (IQR)	105	35 (34-38)	33 (29-36)	33 (29-36)	.12
Education level, n (%)	105				.11
University degree		3 (43)	50 (51)	53 (50.5)	
Primary school		2 (29)	4 (4)	6 (5.7)	
High school		2 (29)	44 (45)	46 (43.8)	
No alcohol, n (%)	105	7 (100)	92 (94)	99 (94.3)	.50
No smoking, n (%)	105	7 (100)	95 (97)	102 (97.1)	.64
Previous pregnancies, n (%)	105				
0		2 (29)	50 (51)	52 (49.5)	.45
≥1		5 (72)	48 (49)	53 (50.4)	
Outcome of pregnancies, n (%)	53				
Birth at term		3 (43)	30 (31)	33 (31)	.50
Premature		1 (14)	2 (2)	3 (3)	.06
Stillbirth		0 (0)	1 (1)	1 (1)	.79
Miscarriage		2 (29)	26 (27)	28 (27)	.90
Extrauterine pregnancy		0 (0)	1 (1)	1 (1)	.79
Hydatidiform mole		0 (0)	0 (0)	0 (0)	
Actual pregnancy problem, n (%)	105				
Gestational hypertension induced, pre-eclampsia, eclampsia		0 (0)	11 (11)	11 (10.5)	.64
Partial placental abruption		1 (14)	1 (1)	2 (1.9)	.01
Placenta previa		0 (0)	4 (4)	4 (3.8)	.59
Breech presentation of the fetus		0 (0)	1 (1)	1 (1.0)	.79
Gestational diabetes		0 (0)	7 (7)	7 (6.7)	.46
Intrahepatic cholestasis of pregnancy		0 (0)	14 (14)	14 (13.3)	.28
Pregnancy hyperemesis		0 (0)	3 (3)	3 (2.9)	.64
Twin pregnancy		1 (14)	15 (15)	16 (15.2)	.94
Risk of premature birth		3 (43)	38 (39)	41 (39.0)	.83
Intrauterine growth restriction		0 (0)	8 (8)	8 (7.6)	.43
Other		3 (43)	26 (27)	29 (27.6)	.66
Tool used to search health information, n (%)	105				
Smartphone		2 (29)	79 (81)	81 (77.1)	.002
Tablet		0 (0)	2 (2)	2 (1.9)	.70
Notebook		1 (14)	32 (33)	33 (31.4)	.31
Personal computer		4 (57)	3 (3)	7 (6.7)	<.001
Last time of internet use for health information, n (%)	95				.85
Within the last week		5 (83)	79 (89)	84 (88)	
Within the last month		1 (17)	9 (10)	10 (11)	
Within the last year		0 (0)	0 (0)	0 (0)	
Over a year ago		0 (0)	1 (1)	1 (1)	

	N	Did not use the internet before hospitalization (n=7)	Used the internet before hospitalization (n=98)	All subjects	*P* value
Usefulness of online information about the pathology, n (%)	95				.09
Very useful		2 (33)	7 (8)	9 (9)	
Somewhat useful		3 (50)	73 (82)	76 (80)	
A little useful		1 (17)	9 (10)	10 (11)	
Not at all useful		0 (0)	0 (0)	0 (0)	
Do not share online health information with midwife or gynecologist, n (%)	93	3 (50)	46 (53)	49 (53)	.89
To what degree do you feel safe consulting the internet for advice or information on pregnancy?, n (%)	95				.84
Completely safe		0 (0)	2 (2)	2 (2)	
Very confident		0 (0)	11 (12)	11 (12)	
Fairly confident		5 (83)	55 (62)	60 (63)	
Shortly confident		1 (17)	20 (22)	21 (22)	
Not at all confident		0 (0)	1 (1)	1 (1)	
STAI[a] trait score, median (IQR)	105	45 (42-47)	47 (45-50)	47 (44-50)	.18
EQ-5D[b] score, median (IQR)	104	0.80 (0.57-0.85)	0.76 (0.69-0.88)	0.76 (0.69-0.88)	.96

[a]STAI: State-Trait Anxiety Inventory.

[b]EQ-5D: EuroQOL 5 dimensions.

Figure 1. Use of Internet Health-information Questionnaire (UIH)-visual analog scale (VIS) levels over the days spent in hospital.

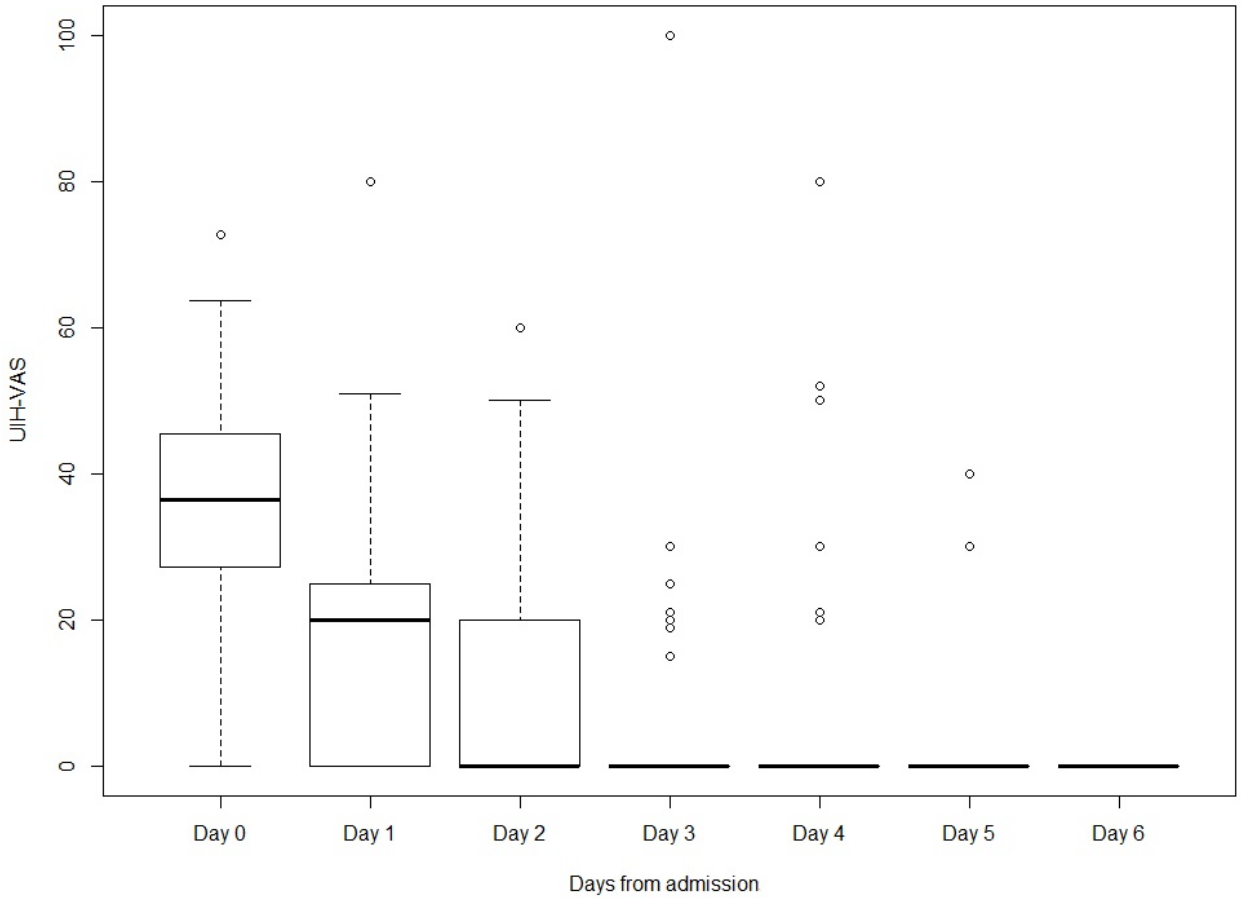


Table 2. Behavior of internet use, health status, critical events, anxiety, and drug use on the first two days of hospitalization and at discharge.

	Day 1				Day 2				Discharge			
	All (N=105)	Internet use (n=81)	No internet use (n=24)	*P* value	All (N=105)	Internet use (n=57)	No internet use (n=48)	*P* value	All (N=105)	Internet use (n=11)	No internet use (n=94)	*P* value
State of health today (VAS[a] scale), median (IQR)	59 (50-70)	50 (50-65)	70 (60-80)	<.001	60 (50-70)[b]	50 (50-70)	70 (50-70)	.004	80 (70-90)[c]	50 (50-50)	80 (70-90)	.004
Critical events that altered your emotional state today, n (%)												
Family	N/A[d]	0 (0)	N/A	N/A	N/A	0 (0)	N/A	N/A	N/A	0 (0)	N/A	N/A
Obstetric	N/A	76 (95)	N/A	N/A	N/A	54 (95)	N/A	N/A	N/A	11 (100)	N/A	N/A
Hospital-related	N/A	3 (4)	N/A	N/A	N/A	2 (4)	N/A	N/A	N/A	0 (0)	N/A	N/A
Other	N/A	1 (1)	N/A	N/A	N/A	(1 (2)	N/A	N/A	N/A	0 (0)	N/A	N/A
UIH[e]-VAS, median (IQR)	20 (0-25)[f]	20 (0-29)	0 (0.00-8.75)	.02	0 (0-20)	20 (0-30)	0 (0-0)	<.001	0 (0-0)	0 (0-0)	0 (0-0)	.84
STAI[g]-state score, median (IQR)	42 (41-44)	42 (41-44)	42 (40-44)	.83	42 (41-44)	42 (40-44)	42 (41-44)	.55	41 (40-43)	41 (38-43)	41 (40-44)	.11
Use of drugs, n (%)	93 (88.6)	73 (90)	20(83)	.36	86 (81.9)	52 (91)	34 (71)	.007	47 (44.8)	7 (64)	40 (43)	.18

[a]VAS: visual analog scale.

[b]N=103.

[c]N=104.

[d]N/A: not applicable; these data were only assessed among subjects that reported using the internet.

[e]UIH: Use of Internet Health-information.

[f]N=81.

[g]STAI: State-Trait Anxiety Inventory.

Anxiety levels were stable over time (Figure 2). Overall, the results indicated that using the web as a source of health information does not substantially increase anxiety levels.

A multivariate model was used to estimate the association between STAI scores and UIH-VAS in the first two days of hospitalization (Table 3). Only the UIH-VAS scale showed a significant nonlinear association (*P*=.007), which remained significant after adjustment for major confounding factors (Figure 3). No significant interaction was found between UIH-VAS and time (day 1, day 2) on STAI (*P*=.51).

Figure 2. Anxiety levels, determined by the State-Trait Anxiety Inventory (STAI) state score during the first 6 days of hospitalization from admission (*P*=.33).

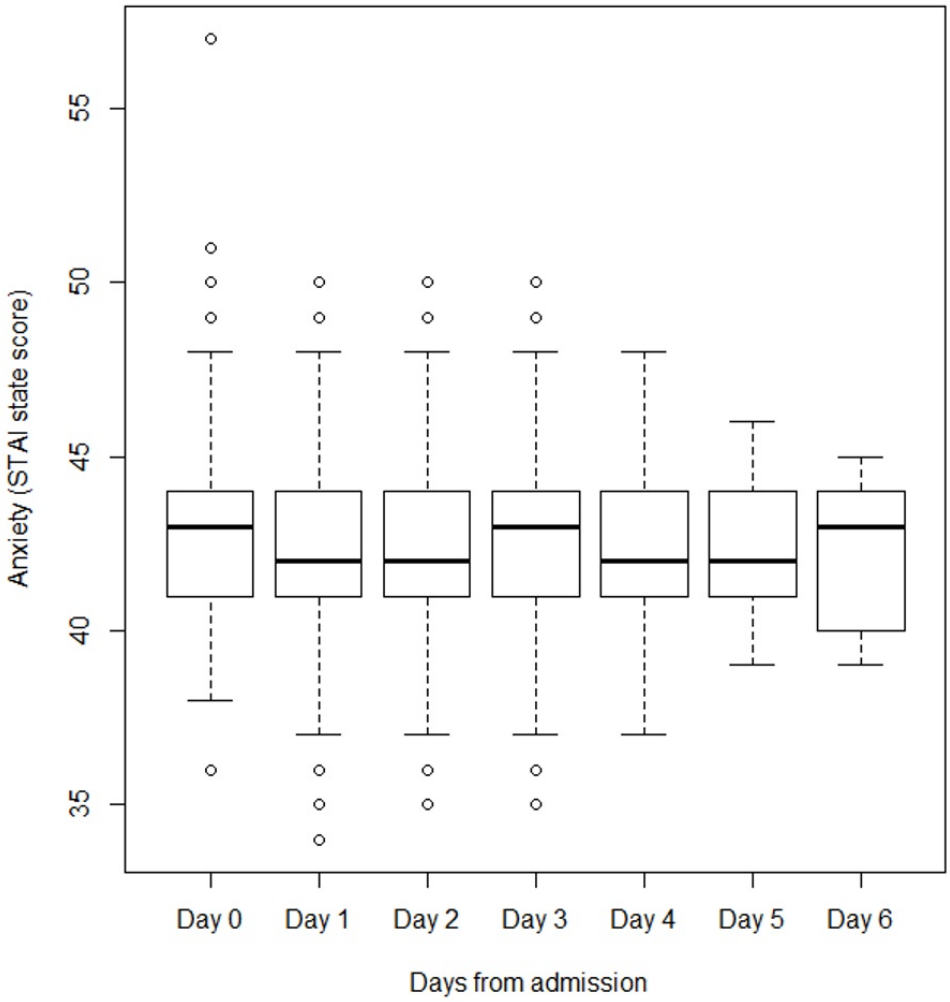


Table 3. Multivariate model for State-Trait Anxiety Inventory score.

Covariate	Effect[a]	SE	Lower 0.95	Upper 0.95	*P* value
Age (7-year difference)	–0.435	0.239	–0.906	0.035	.07
EQ-5D[b]-VAS[c] (0.20 points difference)	0.009	0.145	–0.276	0.296	.95
UIH[d]-VAS (20 points difference after 30 points)	–1.855	0.596	–3.031	–0.680	.007
Drug consumption (no vs yes)	–0.157	0.570	–1.282	0.967	.78
Critical events (occurrence vs nonoccurrence)	–0.444	0.450	–1.332	0.443	.32

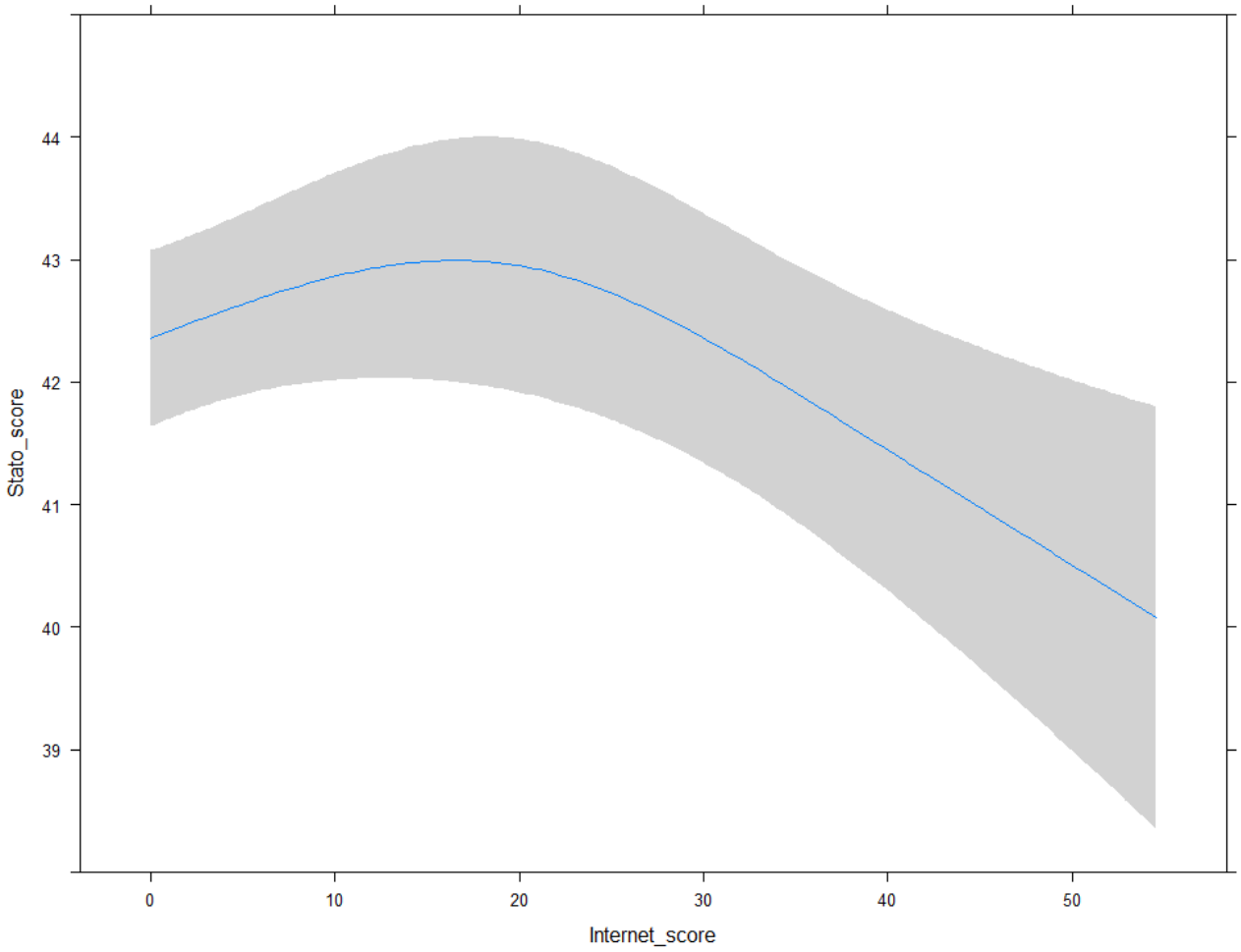
[a]Effect is the slope of the linear regression model for each covariate expressed in terms of the interquartile difference for continuous covariates and using a reference category for categorical variables; for UIH-VAS, the effect is nonlinear.

[b]EQ-5D: EuroQOL 5 dimensions.

[c]VAS: visual analog scale.

[d]UIH: Use of Internet Health-information.

Figure 3. Association of UIH-VAS and STAI-State score. Non linearity (*P*=.007) estimated via restricted cubic splines and adjusted for EQ5D-VAS, age, critical events, and drug consumption. UIH: Use of Internet Health-information Questionnaire; VAS: visual analog scale; STAI: State-Trait Anxiety Inventory; EQ5D: EuroQOL 5 dimensions questionnaire.



Discussion

Principal Findings and Interpretation

The results of the present study need to be interpreted in light of the related literature on network system expansion [29,30]. An increasing number of people are browsing the internet daily to obtain any type of information. Access and usage of the internet is now nearly ubiquitous, which poses new challenges for health care practitioners and users, and the terms "pregnancy" and "obstetrics" are among the top 5 searched medical keywords [31]. In addition, when defining online health communication as sending emails about health matters to family or health care providers [32], 52.4% of the women (55/105) that had internet access in our study reported sending emails or using the internet to communicate with a doctor about their health.

Although we did not find significant associations between factors such as age or education with internet use, this effect partially reflects findings from previous studies [24] as we found a large diffusion of web use among a relatively young sample (median 33 years old), with 93% of the population accessing the internet to obtain nonspecific information about health.

Studies published in the early 2000s indicated moderate use of online health information-seeking by internet users in the general population [33,34]. Conversely, but not surprisingly, despite focusing only on pregnant women in this study, we found a high percentage of women using the internet to search for information about pregnancy problems before hospital admission (82%). Other studies showed that 91% of the surveyed women had access to the internet, 84% of whom used it to search for information related to their condition, especially in the early stages of gestation, whereas 70% of these women did not talk to their health care providers about the health information they found online [35]. Since half of the information sought by the women in our sample was suggested by physicians, the internet was used most likely used to obtain information that could confirm the diagnosis or provide further details on the topic. Nevertheless, the women in our cohort also did not largely discuss what they found with physicians, probably because they felt that their health care providers would not accept the internet as a reliable source of medical information [35]. Finally, patients are usually considered as passive recipients of information rather than being treated as the main actors in their health course, as it should be. This general situation can also be applied to pregnant women who seek support and a sense of community in relation to their condition [11].

Strengths and Limitations

To our knowledge, this is the first study to directly evaluate internet use by pregnant women during hospitalization for

obstetric problems. Although hospitalization causes an increase in anxiety levels in this vulnerable population, our results showed that use of the internet to search health information reduces anxiety levels. The reason behind this finding could be related to the effect of the acquisition of information itself; that is, anxiety (state anxiety) can be reduced when pregnant women become more aware about their clinical condition (ie, the prognosis of the disease and its management). The majority of information received from the internet was obtained in the first 2 days of hospitalization. The reduction in internet usage from the third day of hospitalization is likely due to the longer time spent in contact with physicians, the influence of setting and health care providers, and clinical improvement. Consequently, the information sought during the first few days of hospitalization likely helped the pregnant women in reducing their anxiety levels.

Moreover, pregnant women often receive limited basic information on prenatal health behaviors. Patients often perceive this information to be inconsistent and inadequate, which could also explain why they search for information online and do not share it with their physicians [36]. This suggests that current assistance approaches to pregnancy may not fully respond to patients' information demand. This might be due to the limited time of direct contact between patients and their health care providers but also due to the unpredictability of the onset of disease and the complexity of diagnosis. Moreover, disease management by midwives requires time for both communicating information and understanding it.

Pregnancy disorders have a clear impact on the perception of anxiety; consequently, the risk of adverse events for the mother and her baby imposes some lifestyle changes to a pregnant woman. In this context, information can play an important role on women's psychological status: improved knowledge about a disease will increase a patient's perceived self-efficacy and the ability to develop adaptive coping strategies. The internet offers the possibility to remain connected with the virtual community of pregnant women and physicians and to obtain all types of information, making internet users more confident in how to manage their condition [32,37]. Some authors also hypothesized that people looking for medical advice and health information are more predisposed to pay greater attention to and be more interested in their clinical condition, resulting in a higher self-efficacy perception [33,36].

Interestingly, our study showed a quite substantial potential impact of the internet in reducing anxiety. Patients with higher internet usage behavior reported an anxiety level that was 2 points lower than that of patients with less intense internet usage (42 vs 40 points, P=.008). This effect accounts for approximately one quarter of the effect of more aggressive therapies in reducing pathological anxiety, such as serotonin reuptake inhibitors combined with psychotherapy, and psychotherapy treatment alone, and accounts for approximately one half of the effect of cognitive behavioral therapy and other unconventional therapies [24-39]. Since anxiety is modulated by many intersectional factors, it might be interesting to further evaluate the effect of internet usage in association with other types of treatments for anxiety in pregnant women, even if a diagnosis of pathological anxiety would be necessary and certain antianxiety drugs cannot be administered to pregnant women.

This study also has several limitations. First, the UIH questionnaire, despite being validated, is not very detailed in terms of assessing the quality of internet usage. Furthermore, data on the specific websites visited would have provided a more precise framework of internet usage. This might be particularly important in assessing the quality of the obtained information about the disease and its subsequent impact on anxiety and other forms of psychological distress. Moreover, the names and types of websites would have been useful to investigate the emotional status in relation to active (eg, sharing health information with others) and passive (eg, simple search for information for personal purposes) use of the internet.

Finally, because of the nature of this study, causal interpretation of the association between exposure to the internet and the level of anxiety is not possible, making the potential interpretation on the "therapeutic" psychological effect of internet usage merely speculative at this point.

Conclusions

This study has implications for health care providers, suggesting that the internet could offer a useful instrument to support clinical practice due to its informational power and its potential impact on well being–related outcomes. The widespread search for online health information among women with pregnancy-related diseases mainly focuses on the possible outcomes for the baby and on the quality of communication between patients and health care providers, emphasizing the role of the internet as a potential tool for enhancement of such essential communication.

To effectively influence the online experiences of pregnant women, professionals involved in the childbirth pathway should have a basic understanding of the internet and learn how to actively engage women's interest in the internet. For this purpose, the installation of free wifi areas in maternity departments could be useful.

Acknowledgments

We are grateful to all of the midwives of the obstetrics departments of the two hospitals for helping with recruitment for this study, and we thank the pregnant women who agreed to participate in the study.

Conflicts of Interest

None declared.

References

1. Lumley J. Defining the problem: the epidemiology of preterm birth. BJOG 2003 Apr;110(Suppl 20):3-7 [FREE Full text] [Medline: 12763104]
2. Cumberbatch C, Birndorf C, Dresner N. Psychological implications of high-risk pregnancy. Int J Fertil Womens Med 2005;50(4):180-186. [Medline: 16405103]
3. Brandon AR, Trivedi MH, Hynan LS, Miltenberger PD, Labat DB, Rifkin JB, et al. Prenatal depression in women hospitalized for obstetric risk. J Clin Psychiatry 2008 Apr;69(4):635-643 [FREE Full text] [doi: 10.4088/jcp.v69n0417] [Medline: 18312059]
4. Ibanez G, Charles M, Forhan A, Magnin G, Thiebaugeorges O, Kaminski M, EDEN Mother–Child Cohort Study Group. Depression and anxiety in women during pregnancy and neonatal outcome: data from the EDEN mother-child cohort. Early Hum Dev 2012 Aug;88(8):643-649. [doi: 10.1016/j.earlhumdev.2012.01.014] [Medline: 22361259]
5. World Health Organization. The global burden of disease: 2004 update. URL: http://www.who.int/healthinfo/global_burden_disease/2004_report_update/en/ [accessed 2019-04-23]
6. Cripe SM, Frederick IO, Qiu C, Williams MA. Risk of preterm delivery and hypertensive disorders of pregnancy in relation to maternal co-morbid mood and migraine disorders during pregnancy. Paediatr Perinat Epidemiol 2011 Mar;25(2):116-123 [FREE Full text] [doi: 10.1111/j.1365-3016.2010.01182.x] [Medline: 21281324]
7. Grote NK, Bridge JA, Gavin AR, Melville JL, Iyengar S, Katon WJ. A meta-analysis of depression during pregnancy and the risk of preterm birth, low birth weight, and intrauterine growth restriction. Arch Gen Psychiatry 2010 Oct 04;67(10):1012-1024 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2010.111] [Medline: 20921117]
8. Bodnar LM, Wisner KL, Moses-Kolko E, Sit DKY, Hanusa BH. Prepregnancy body mass index, gestational weight gain, and the likelihood of major depressive disorder during pregnancy. J Clin Psychiatry 2009 Sep;70(9):1290-1296 [FREE Full text] [doi: 10.4088/JCP.08m04651] [Medline: 19607761]
9. Borri C, Mauri M, Oppo A, Banti S, Rambelli C, Ramacciotti D, et al. Axis I psychopathology and functional impairment at the third month of pregnancy: Results from the Perinatal Depression-Research and Screening Unit (PND-ReScU) study. J Clin Psychiatry 2008 Oct;69(10):1617-1624. [doi: 10.4088/jcp.v69n1012] [Medline: 19192445]
10. Byatt N, Hicks-Courant K, Davidson A, Levesque R, Mick E, Allison J, et al. Depression and anxiety among high-risk obstetric inpatients. Gen Hosp Psychiatry 2014;36(6):644-649 [FREE Full text] [doi: 10.1016/j.genhosppsych.2014.07.011] [Medline: 25149040]
11. Diaz JA, Griffith RA, Ng JJ, Reinert SE, Friedmann PD, Moulton AW. Patients' use of the Internet for medical information. J Gen Intern Med 2002 Mar;17(3):180-185 [FREE Full text] [doi: 10.1046/j.1525-1497.2002.10603.x] [Medline: 11929503]
12. Lagan BM, Sinclair M, Kernohan WG. A Web-based survey of midwives' perceptions of women using the Internet in pregnancy: a global phenomenon. Midwifery 2011 Apr;27(2):273-281. [doi: 10.1016/j.midw.2009.07.002] [Medline: 19700228]
13. Usui N, Kamiyama M, Tani G, Kanagawa T, Fukuzawa M. Use of the medical information on the internet by pregnant patients with a prenatal diagnosis of neonatal disease requiring surgery. Pediatr Surg Int 2011 Dec 11;27(12):1289-1293. [doi: 10.1007/s00383-011-2965-6] [Medline: 21833721]
14. Cline RJ, Haynes KM. Consumer health information seeking on the Internet: the state of the art. Health Educ Res 2001 Dec;16(6):671-692. [doi: 10.1093/her/16.6.671] [Medline: 11780707]
15. Sacks S, Abenhaim HA. How evidence-based is the information on the internet about nausea and vomiting of pregnancy? J Obstet Gynaecol Can 2013 Aug;35(8):697-703. [doi: 10.1016/S1701-2163(15)30859-8] [Medline: 24007704]
16. Norr AM, Capron DW, Schmidt NB. Medical information seeking: impact on risk for anxiety psychopathology. J Behav Ther Exp Psychiatry 2014 Sep;45(3):402-407. [doi: 10.1016/j.jbtep.2014.04.003] [Medline: 24818986]
17. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009 Apr;42(2):377-381 [FREE Full text] [doi: 10.1016/j.jbi.2008.08.010] [Medline: 18929686]
18. Chou WS, Hunt YM, Beckjord EB, Moser RP, Hesse BW. Social media use in the United States: implications for health communication. J Med Internet Res 2009;11(4):e48 [FREE Full text] [doi: 10.2196/jmir.1249] [Medline: 19945947]
19. AlGhamdi KM, Moussa NA. Internet use by the public to search for health-related information. Int J Med Inform 2012 Jun;81(6):363-373. [doi: 10.1016/j.ijmedinf.2011.12.004] [Medline: 22217800]
20. Hallila LE. Nursing students' use of Internet and Computer for their Education in the College of Nursing. Int J Nurs Clin Pract 2014 Dec 25;1(1):IJNCP-108. [doi: 10.15344/2394-4978/2014/108]
21. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736 [FREE Full text] [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
22. Spielberg CD. State-trait anxiety inventory: A comprehensive bibliography. Palo Alto, CA: Consulting Psychologists Press; 1984.
23. Spielberg CD. Inventario per l'ansia di. Firenze: Organizzazioni Speciali; 1989.
24. Minto C, Bauce B, Calore C, Rigato I, Folino F, Soriani N, et al. Is Internet use associated with anxiety in patients with and at risk for cardiomyopathy? Am Heart J 2015 Jul;170(1):87-95. [doi: 10.1016/j.ahj.2015.02.024] [Medline: 26093868]

25. Huber PJ. The behavior of maximum likelihood estimates under nonstandard conditions. 1967 Presented at: Fifth Berkeley Symposium on Mathematical Statistics and Probability; 1967; Berkeley, CA.
26. Boisbunon A, Canu S, Fourdrinier D, Strawderman W, Wells MT. Akaike's Information Criterion, Cp and Estimators of Loss for Elliptically Symmetric Distributions. Int Stat Rev 2014 Aug 18;82(3):422-439. [doi: 10.1111/insr.12052]
27. Harrel FE. Regression Modeling Strategies. Berlin: Springer International Publishing; 2015.
28. R Core Team. R: a language and environment for statistical computing. 2017. URL: https://www.R-project.org/ [accessed 2019-04-20]
29. Ekman A, Litton J. New times, new needs; e-epidemiology. Eur J Epidemiol 2007;22(5):285-292. [doi: 10.1007/s10654-007-9119-0] [Medline: 17505896]
30. National Telecommunications and Information Administration. Falling through the net II: new data on the digital divide. URL: https://www.ntia.doc.gov/report/1998/falling-through-net-ii-new-data-digital-divide [accessed 2019-04-23]
31. Spink A, Yang Y, Jansen J, Nykanen P, Lorence DP, Ozmutlu S, et al. A study of medical and health queries to web search engines. Health Info Libr J 2004 Mar;21(1):44-51. [doi: 10.1111/j.1471-1842.2004.00481.x] [Medline: 15023208]
32. Romano AM. A Changing Landscape: Implications of Pregnant Women's Internet Use for Childbirth Educators. J Perinat Educ 2007;16(4):18-24 [FREE Full text] [doi: 10.1624/105812407X244903] [Medline: 18769519]
33. Cotten SR, Gupta SS. Characteristics of online and offline health information seekers and factors that discriminate between them. Soc Sci Med 2004 Nov;59(9):1795-1806. [doi: 10.1016/j.socscimed.2004.02.020] [Medline: 15312915]
34. Morahan-Martin JM. How internet users find, evaluate, and use online health information: a cross-cultural review. Cyberpsychol Behav 2004 Oct;7(5):497-510. [doi: 10.1089/cpb.2004.7.497] [Medline: 15667044]
35. Larsson M. A descriptive study of the use of the Internet by women seeking pregnancy-related information. Midwifery 2009 Feb;25(1):14-20. [doi: 10.1016/j.midw.2007.01.010] [Medline: 17408822]
36. Leiferman J, Sinatra E, Huberty J. Pregnant women's perceptions of patient-provider communication for health behavior change during pregnancy. Open J Obstet Gynecol 2014;04(11):672-684. [doi: 10.4236/ojog.2014.411094]
37. Jacobson P. Empowering the physician-patient relationship: The effect of the Internet. Partnership 2007 May 29;2(1). [doi: 10.21083/partnership.v2i1.244]
38. Marchesi C, Ossola P, Amerio A, Daniel BD, Tonna M, De Panfilis C. Clinical management of perinatal anxiety disorders: A systematic review. J Affect Disord 2016 Jan 15;190:543-550. [doi: 10.1016/j.jad.2015.11.004] [Medline: 26571104]
39. Ross LE, McLean LM. Anxiety disorders during pregnancy and the postpartum period: A systematic review. J Clin Psychiatry 2006 Aug;67(8):1285-1298. [doi: 10.4088/jcp.v67n0818] [Medline: 16965210]

Abbreviations

EQ-5D: European quality of life 5 dimensions questionnaire
STAI: State-Trait Anxiety Inventory
UIH: Use of Internet Health-information Questionnaire
VAS: visual analogic scale

Edited by C Lovis; submitted 28.10.19; peer-reviewed by P Berchialla, A Linn; comments to author 01.02.20; revised version received 16.02.20; accepted 20.02.20; published 06.05.20.

Please cite as:
Coglianese F, Beltrame Vriz G, Soriani N, Piras GN, Comoretto RI, Clemente L, Fasan J, Cristiano L, Schiavinato V, Adamo V, Marchesoni D, Gregori D
Effect of Online Health Information Seeking on Anxiety in Hospitalized Pregnant Women: Cohort Study
JMIR Med Inform 2020;8(5):e16793
URL: https://medinform.jmir.org/2020/5/e16793
doi:10.2196/16793
PMID:32374268

Original Paper

Evaluation of the Quadri-Planes Method in Computer-Aided Diagnosis of Breast Lesions by Ultrasonography: Prospective Single-Center Study

Liang Yongping[1*], MD; Zhang Juan[1*], MD; Ping Zhou[1], MD; Zhao Yongfeng[1], MD; Wengang Liu[1], MD; Yifan Shi[1], MD

The Xiangya Medical School, Central South University, Changsha, Hunan, China

[*]these authors contributed equally

Corresponding Author:
Ping Zhou, MD
The Xiangya Medical School
Central South University
172 Laodong W Rd
Tianxin District
Changsha, Hunan, 410015
China
Phone: 86 731 88618403
Fax: 86 731 88618403
Email: zhouping1000@hotmail.com

Abstract

Background: Computer-aided diagnosis (CAD) is a tool that can help radiologists diagnose breast lesions by ultrasonography. Previous studies have demonstrated that CAD can help reduce the incidence of missed diagnoses by radiologists. However, the optimal method to apply CAD to breast lesions using diagnostic planes has not been assessed.

Objective: The aim of this study was to compare the performance of radiologists with different levels of experience when using CAD with the quadri-planes method to detect breast tumors.

Methods: From November 2018 to October 2019, we enrolled patients in the study who had a breast mass as their most prominent symptom. We assigned 2 ultrasound radiologists (with 1 and 5 years of experience, respectively) to read breast ultrasonography images without CAD and then to perform a second reading while applying CAD with the quadri-planes method. We then compared the diagnostic performance of the readers for the 2 readings (without and with CAD). The McNemar test for paired data was used for statistical analysis.

Results: A total of 331 patients were included in this study (mean age 43.88 years, range 17-70, SD 12.10), including 512 lesions (mean diameter 1.85 centimeters, SD 1.19; range 0.26-9.5); 200/512 (39.1%) were malignant, and 312/512 (60.9%) were benign. For CAD, the area under the receiver operating characteristic curve (AUC) improved significantly from 0.76 (95% CI 0.71-0.79) with the cross-planes method to 0.84 (95% CI 0.80-0.88; $P<.001$) with the quadri-planes method. For the novice reader, the AUC significantly improved from 0.73 (95% CI 0.69-0.78) for the without-CAD mode to 0.83 (95% CI 0.80-0.87; $P<.001$) for the combined-CAD mode with the quadri-planes method. For the experienced reader, the AUC improved from 0.85 (95% CI 0.81-0.88) to 0.87 (95% CI 0.84-0.91; $P=.15$). The kappa indicating consistency between the experienced reader and the novice reader for the combined-CAD mode was 0.63. For the novice reader, the sensitivity significantly improved from 60.0% for the without-CAD mode to 79.0% for the combined-CAD mode ($P=.004$). The specificity, negative predictive value, positive predictive value, and accuracy improved from 84.9% to 87.8% ($P=.53$), 76.8% to 86.7% ($P=.07$), 71.9% to 80.6% ($P=.13$), and 75.2% to 84.4% ($P=.12$), respectively. For the experienced reader, the sensitivity improved significantly from 76.0% for the without-CAD mode to 87.0% for the combined-CAD mode ($P=.045$). The NPV and accuracy moderately improved from 85.8% and 86.3% to 91.0% ($P=.27$) and 87.0% ($P=.84$), respectively. The specificity and positive predictive value decreased from 87.4% to 81.3% ($P=.25$) and from 87.2% to 93.0% ($P=.16$), respectively.

Conclusions: S-Detect is a feasible diagnostic tool that can improve the sensitivity, accuracy, and AUC of the quadri-planes method for both novice and experienced readers while also improving the specificity for the novice reader. It demonstrates important application value in the clinical diagnosis of breast cancer.

Trial Registration: ChiCTR.org.cn 1800019649; http://www.chictr.org.cn/showproj.aspx?proj=33094

(JMIR Med Inform 2020;8(5):e18251) doi:10.2196/18251

KEYWORDS

ultrasonography; breast neoplasm; breast imaging reporting and data system (bi-rads); breast neoplasm diagnosis; cancer screening; computer-aided diagnosis; breast cancer

Introduction

Breast cancer is one of the most common cancers in women and the second leading cause of cancer-related mortality worldwide [1,2]. Early diagnosis of breast cancer can increase the treatment options and survival rate of patients [3]; however, early diagnosis depends on accurate and reliable diagnosis using medical imageology. As a convenient modality, breast ultrasonography plays an important role in breast cancer screening. Despite the improvements in ultrasound diagnosis with the application of new technology, dependence on operator experience remains the main limitation of ultrasound-based diagnosis [4,5]. S-Detect is a recently developed computer-aided diagnosis (CAD) system for breast cancer that provides assistance in morphological analysis based on the Breast Imaging Reporting and Data System (BI-RADS) lexicon and classification [6]. Many studies have reported that the S-Detect system has potential to become a novel diagnostic tool for radiologists [7-10].

In our previous study, the sensitivity was too high in the cross-planes method because it considered the lesion to be malignant if any image of 2 planes indicated malignancy, leading to a decrease in specificity. No study has evaluated the diagnostic performance of CAD in breast lesions with respect to diagnostic planes (cross-plane and quadri-plane methods). Therefore, the purpose of this study was to compare the performance of radiologists with different levels of experience in detecting breast cancer using CAD with the quadri-planes method.

Methods

Patient Selection

We prospectively enrolled patients in our study from November 2018 to October 2019. All patients underwent grayscale breast ultrasound examination before surgery. All lesions were examined after surgery to confirm their pathological type. This prospective single center study was approved by the Institutional Review Board of Third Xiangya Hospital. Informed consent was obtained from all patients.

The inclusion criteria were as follows: patients aged 17-70 years with breast tumors requiring surgery. The exclusion criteria were a history of neoadjuvant chemotherapy or endocrine therapy before surgery, lesions punctured by core-needle biopsy or a Mammotome system, equipment of the breast with a prosthesis, unclear lesions as displayed by ultrasound images, and unwillingness to take part in the study.

Ultrasound Image Acquisition

All images were obtained with an RS80A ultrasound system (Samsung Medison Co Ltd) with a 5-13 megahertz bandwidth (8.4 MHz center frequency) linear transducer. All ultrasound examinations were performed by an independent radiologist with 5 years of experience. In the cross-planes method, 2 typical images of the tumor in the longitudinal and transverse planes were stored in the ultrasound system; in the quadri-planes method, 2 additional cross-plane images were acquired by rotating the probe 45 degrees around the center of the mass.

Computer-Aided Diagnostic System

Our CAD system, S-Detect, extracts features using an integration of artificial neural network classifiers internally installed in the RS80A ultrasound equipment. The sensitivity of the instrument was set to the default. To test the reproducibility of the CAD marks with the same image, we randomly selected 20/512 (3.9%) examinations and passed them through the CAD system 3 times; the results showed that the markings were consistent in all images.

In S-Detect, the cursor was placed on the identified center of the lesion, and a region of interest was automatically drawn along the border of the mass by the ultrasound system. If the borderline was considered inaccurate in any area of the tumor, it was manually edited to achieve the optimum fitness. The ultrasound image features of the lesion were analyzed according to the BI-RADS lexicon, and the final classifications were automatically performed by the ultrasound system. In the S-Detect system, the final assessment classification was divided into dichotomous results of “possibly benign” or “possibly malignant.”

Diagnostic Criteria

According to the fifth version of BI-RADS, the radiologists classified the lesions from category 3 to category 5. BI-RADS category 4 was further subdivided into categories 4A, 4B, and 4C. Category 3 is considered probably benign (<2% likelihood of malignancy), and categories 4A, 4B, and 4C range from low to high suspicion (2%-10%, 10%-50%, and 50%-95% likelihood of malignancy, respectively). Category 5 indicates a high malignancy rate (>95% likelihood of malignancy). Malignant signs in breast ultrasound imaging include irregular shape, antiparallel orientation, noncircumscribed margin, microcalcification, acoustic halo, posterior shadowing, and abnormalities of the surrounding tissue. Lesions with no definitive malignant sign were assigned to category 3; lesions with 1, 2, and 3 malignant signs were assigned to categories 4A, 4B, and 4C, respectively; and lesions with more than 4 malignant signs were assigned to category 5. Accordingly, category 3 and 4A lesions were regarded as benign, and category 4B, 4C, and 5 lesions were regarded as malignant [11,12].

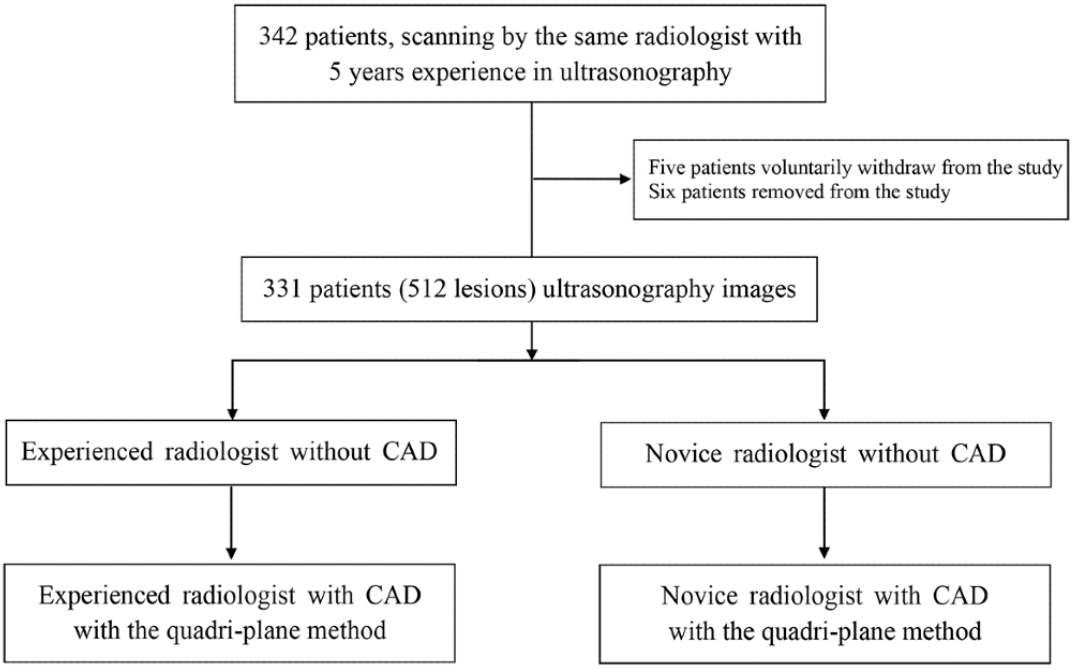
To assess the combination of ultrasound and the CAD system, we acquired images of the longitudinal and transverse planes of the tumor for CAD with the cross-planes method. If 1 plane indicated "possibly malignant," the outcome was considered positive, and the BI-RADS category diagnosis was increased by 1 level (ie, 3 to 4A, 4A to 4B, 4B to 4C, 4C to 5). If both planes indicated "possibly benign," the outcome was considered negative, and the BI-RADS category diagnosis was decreased by 1 level (ie, 5 to 4C, 4C to 4B, 4B to 4A, 4A to 3) [13]. For the quadri-planes method, if any 2 planes indicated "possibly malignant," the outcome was considered positive, and the BI-RADS category diagnosis was increased by 1 level. If all 4 planes indicated "possibly benign," the outcome was considered negative, and the BI-RADS category diagnosis was decreased by 1 level.

Readers, Reading Modes, and Training

The study included 2 readers: a novice reader with 1 year of ultrasound experience and an experienced reader with 5 years of ultrasound experience. Both readers were trained in the reading procedures with 20 ultrasound images (from the 512 examinations) that were not part of the study set, 10 of which were read without using CAD (without-CAD mode). The readers assessed the other 10 images in combined-CAD mode with the cross-planes method and the quadri-planes method; the readers first read the ultrasound images without using CAD and then mechanically combined the indications of the CAD marks to make the final decision.

Both readers performed every examination in each reading mode independently and were blinded to any information about the patients, including age, manifestation of symptoms, and previous radiology reports. The readers were asked to read for at least 2 hours per day to simulate the typical process of batch reading in such examinations (Figure 1).

Figure 1. The study design and workflow. CAD: computer-aided diagnosis.



Statistical Analysis

Statistical evaluation was performed using SPSS software version 19.0 (IBM Corporation). Taking the pathology results as the gold standard, we analyzed the diagnostic sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC) in without-CAD mode and combined-CAD mode (with the quadri-planes method). The confirmatory diagnosis was defined as the diagnosis made on the basis of pathology. The diagnostic parameters of the combined-CAD mode and without-CAD mode were compared using the McNemar test for sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and accuracy for match-paired data. We used the Hanley and McNeil method to analyze the differences between pairs of AUCs. The number of malignant planes of each tumor was recorded with the quadri-planes method, and the ROC curves were drawn based on the pathological results to determine the cutoff value based on the maximum Youden index. The degree of agreement between the experienced reader in without-CAD mode and the novice reader in combined-CAD mode was analyzed using kappa statistics. The criteria for the kappa values were poor ≤0.2, fair 0.21-0.4, moderate 0.41-0.6, good 0.61-0.8, and perfect 0.81-1 [14]. For all tests mentioned, $P<.05$ was considered to indicate statistical significance.

Results

Characteristics of the Patients and Lesions

The patient demographics and lesion characteristics are summarized in Table 1. A total of 331 patients who presented with 512 lesions were included in this study. The mean age of the examined patients was 43.88 years, range 17-70 (SD 12.10). The diameters of the lesions ranged from 0.26-9.50 centimeters,

mean 1.85 (SD 1.19). Among the 512 breast lesions, 200/512 (39.1%) were malignant and 312/512 (61.9%) were benign. The mean sizes of all lesions were similar and were close to 2 cm; benign lesions were the smallest (1.82 cm) and malignant lesions were the largest (2.28 cm).

Table 1. Patient demographics (N=331) and lesion characteristics (N=512).

Characteristic	Value
Age (years)	
Mean (SD)	43.88 (12.10)
Median (range)	45 (17-70)
Age distribution (years), n (%)	
<30	39 (11.8)
30-39	76 (23.0)
40-49	104 (31.4)
50-59	80 (24.2)
60-70	32 (9.7)
Lesion size (cm)[a]	
Mean (SD)	1.85 (1.19)
Median (range)	1.7 (0.26-9.50)
Malignant lesion size (cm)	
Mean (SD)	2.28 (1.10)
Median (range)	2.18 (0.26-6.20)
Benign lesion size (cm)	
Mean (SD)	1.82 (1.39)
Median (range)	1.41 (0.37-9.50)
Pathological finding, n (%)	
Malignant	200 (39.1)
Benign	312 (60.9)
Histological type, n (%)	
Malignant (n=200)	
Intraductal carcinoma in situ	9 (4.5)
Invasive lobular carcinoma	17 (8.5)
Mucinous adenocarcinoma	4 (2.0)
Medullary carcinoma	3 (1.5)
Invasive ductal carcinoma	167 (83.5)
Benign (n=312)	
Intraductal papilloma	37 (11.9)
Granulomatous mastitis	8 (2.6)
Fibroma	211 (67.6)
Hyperplasia-induced lesions	5 (17.31)
Scar tissue	2 (0.6)

[a]cm: centimeters.

Reader Performance

The diagnostic performance of CAD and of the novice and experienced readers in comparison with the pathological diagnoses is depicted in Table 2.

The statistical evaluation of the performance of the CAD system and of the readers is shown in Table 2. For CAD, the AUC improved significantly between the cross-planes method and the quadri-planes method (Z=4.42, P<.001). The cutoff value of the positive planes in the quadri-planes method was 2.5 based

on the Youden index of 0.68. Considering that breast cancer often demonstrates invasive characteristics and has relatively poor prognosis [15], we set the threshold to any 2 positive planes of 4 images.

For the novice reader, the improvement in the AUCs was significant between the without-CAD mode and combined-CAD mode with the quadri-planes method (Z=5.55, P<.001). However, there was no significant difference in the AUCs for the without-CAD and combined-CAD modes for the experienced reader (Z=1.44, P=.15; Table 3, Figure 2). The kappa indicating consistency between the experienced reader in without-CAD mode and the novice reader in combined-CAD mode was 0.63.

When a BI-RADS category 4A threshold was used, in contrast to CAD with the cross-planes method, significant improvements in specificity (P<.001), PPV (P=.01), and accuracy (P=.03) were observed for the quadri-planes method; however, there was no significant difference in NPV, and the sensitivity decreased. The sensitivity, NPV, and accuracy improved in the combined-CAD mode compared with the without-CAD mode for both readers (Table 3). Among these, the sensitivity improved significantly between the 2 reading modes for both the novice reader (P=.004) and the experienced reader (P=.045), whereas the accuracy improved significantly only for the novice reader. Moreover, there were no significant differences between modes for either reader with respect to specificity, PPV, or NPV.

Table 2. Diagnostic performance of the computer-aided diagnosis system and the novice and experienced readers in the 2 reading modes with the Breast Imaging Reporting and Data System Category 4A threshold. The pathological diagnosis was considered to be the gold standard.

Pathological diagnosis	CAD[a]				Novice reader				Experienced reader			
	Cross-planes method		Quadri-planes method		Without-CAD mode		Combined-CAD mode with quadri-planes		Without-CAD mode		Combined-CAD mode with quadri-planes	
	+[b]	–[c]	+	–	+	–	+	–	+	–	+	–
+	190	10	175	25	120	80	158	42	152	48	174	26
–	137	175	58	254	47	265	38	274	22	290	40	272

[a]CAD: computer-aided diagnosis.

[b]+: positive diagnosis. Breast Imaging Reporting and Data System assessment categories 4B, 4C, and 5 were considered positive for cancer.

[c]–: negative diagnosis.

Table 3. Statistical evaluation of the performance of the computer-aided diagnosis system and the 2 readers with P values indicting differences between various groups.

Characteristic	CAD[a]		Novice reader		Experienced reader		Significance				
	CP[b] method	QP[c] method	Without-CAD mode	Combined-CAD mode with QP	Without-CAD mode	Combined-CAD mode with QP	P value[d]	P value[e]	P value[f]	P value[g]	P value[h]
Sensitivity, %	95.0	87.5	60.0	79.0	76.0	87.0	.048	.004	.045	.61	.04
Specificity, %	56.1	81.4	84.9	87.8	93.0	87.2	<.001	.53	.15	.23	.01
PPV[i], %	58.1	75.1	71.9	80.6	87.4	81.3	.01	.13	.25	.25	.03
NPV[j], %	94.6	91.0	76.8	86.7	85.8	91.3	.27	.07	.27	.84	.27
Accuracy, %	71.3	83.8	75.2	84.4	86.3	87.1	.03	.03	.32	.69	.69
AUC[k]	0.76	0.84	0.73	0.83	0.85	0.87	<.001	<.001	0.15	.58	.76

[a]CAD: computer-aided diagnosis.

[b]CP: cross-planes.

[c]QP: quarter-planes.

[d]P for CAD with the cross-planes method vs CAD with the quadri-planes method.

[e]P for the novice reader without CAD vs the novice reader using CAD with the quadri-planes method.

[f]P for the experienced reader without CAD vs the experienced reader using CAD with the quadri-planes method.

[g]P for the novice reader using CAD with the quadri-planes method vs the experienced reader without CAD.

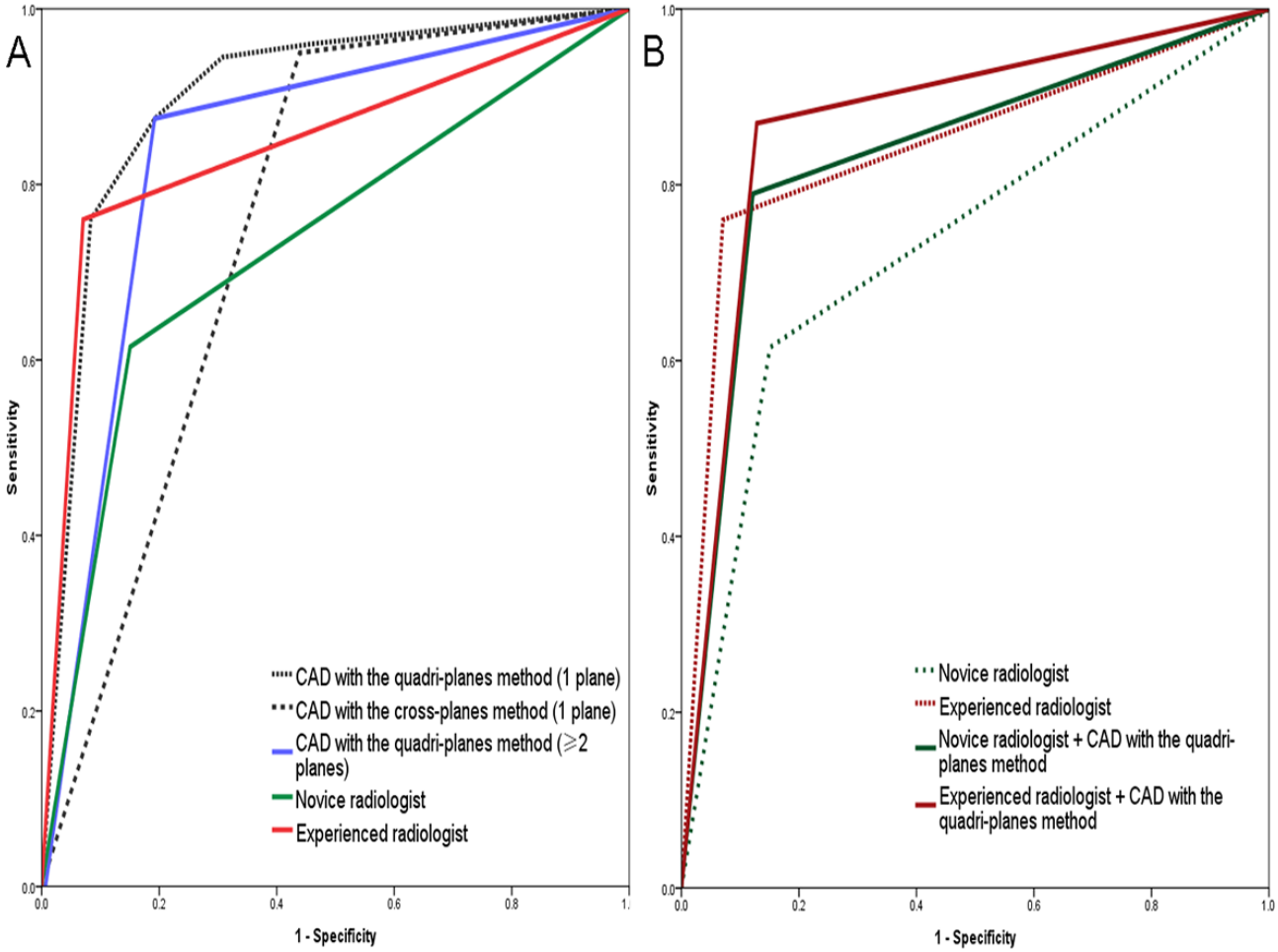
[h]P for CAD with the quadri-planes method vs the experienced reader without CAD.

[i]PPV: positive predictive value.

[i]NPV: negative predictive value.

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

Figure 2. Receiver operating characteristic curves for the CAD method and the readers. (A) CAD with the quadri-plane and cross-plane methods (if either plane was malignant, the result was regarded as positive); CAD with the quadri-planes method with 2 planes as the threshold (if ≥2 planes were malignant, the result was regarded as positive); the novice reader without CAD; the experienced reader without CAD. (B) The novice reader without CAD; the experienced reader without CAD; the novice reader combined with CAD with the quadri-planes method; the experienced reader combined with CAD with the quadri-planes method. AUC: area under the curve; CAD: computer-aided diagnosis.



Management of Diagnostic Decision Changes

In contrast to the mode without CAD, changes in the diagnostic decision with combined-CAD mode were moderately more common for the novice reader than for the experienced reader (115/512, 22.5% vs 70/512, 13.7%; *P*=.09). The proportions of malignancy lesions correctly upgraded from category 4A to 4B by the novice reader and experienced reader were similar (44/115, 38% vs 27/70, 39%; *P*=.88). However, the proportion of benign lesions correctly downgraded from category 4B to 4A by the novice reader was higher than that by experienced reader, with a very close to significant statistical difference (37/115, 32% vs 10/70, 14%; *P*=.05). The proportions of malignancy lesions incorrectly downgraded from category 4B to 4A by both readers were similar (6/115, 5% vs 5/70, 7%). In addition, the proportion of benign lesions incorrectly upgraded by the novice reader was lower than that by the experienced reader (28/115, 24% vs 28/70, 40%). The management decision changes of the 2 readers are provided in Table 4.

The kappa indicating consistency between the experienced reader and the novice reader with combined-CAD mode was slightly higher than that for without-CAD mode (0.63 vs 0.57). These results are provided in Table 5.

Table 4. Management decision changes by the 2 readers using CAD with the quadri-planes method.

Decision change with CAD	Novice reader (n=115)		Experienced reader (n=70)		*P* value	
	Correct, n (%)	Incorrect, n (%)	Correct, n (%)	Incorrect, n (%)	Correct	Incorrect
4A to 4B	44 (38)	28 (24)	27 (39)	28 (40)	.88	<.001
4B to 4A	37 (32)	6 (5)	10 (14)	5 (7)	.05	.47

Table 5. Comparisons of consistency between the experienced reader in without-CAD mode and the novice reader in without-CAD mode and combined-CAD mode with the quadri-planes method.

Experienced reader in without-CAD mode	Novice reader in without-CAD[a] mode[b]		Novice reader in combined-CAD mode[c]	
	+[d]	–[e]	+	–
+	121	53	141	33
–	46	292	55	283

[a]CAD: computer-aided diagnosis.

[b]kappa=0.57.

[c]kappa=0.63.

[d]Positive diagnosis. Breast Imaging Reporting and Data System assessment categories 4B, 4C, and 5 were considered positive for cancer.

[e]Negative diagnosis.

Discussion

Principal Findings

In our study, the AUCs of CAD with the quadri-planes method were significantly higher than those of CAD with the cross-planes method (P<.001); even when we chose any 2 malignant planes as the threshold, the AUC of the quadri-planes method was still higher than that of the cross-planes method (P<.001). The sensitivity, accuracy, and AUC improved for both the novice and experienced readers using combined-CAD mode with the quadri-planes method. Additionally, compared to without-CAD mode, the consistency level improved from fair to good between the novice reader in combined-CAD mode and the experienced reader in without-CAD mode.

Choi et al [10] recently reported that the specificity and AUC of both experienced and inexperienced readers improved using a CAD system combined with S-Detect; moreover, the sensitivity of the inexperienced reader improved significantly. Although the diagnosis performance of the readers improved in Choi's study, the sensitivity of the readers combined with CAD was not satisfactory for detecting breast cancer (66.7% and 75.0%, respectively). These results may have been obtained because the proportion of malignant lesions was too low (6%); moreover, the data in that study were derived from a small number of patients. All these factors can lead to increased false negative results. According to our previous study [16], high sensitivity is a remarkable characteristic of S-Detect in the cross-planes method; this is similar to some previously published studies, where the sensitivity of the ultrasound CAD system was reported to be high (between 88.9% and 100%) [17,18].

It is known that high sensitivity in diagnostic performance can lead to unnecessary breast biopsies and increased medical costs borne by patients; therefore, we developed the quadri-planes method to address this problem. CAD in the quadri-planes method resulted in both improved sensitivity (60.0% to 79.0%) and specificity (84.9% to 87.8%) for the novice reader; in addition, there was no statistically significant change in specificity for the experienced reader (93.0% to 87.2%), while the sensitivity improved significantly (76.0% to 87.0%). This indicates that CAD with the quadri-planes method can improve the sensitivity and specificity of the results reported by readers, especially less experienced readers.

In our investigation, the specificity, accuracy, and AUC of CAD with the quadri-planes method were all higher than those of CAD with the cross-planes method, although the sensitivity of the quadri-plane method was slightly lower. This is likely because the quadri-planes method is based on the cross-planes method; therefore, 2 of the 4 planes in the quadri-planes method were the same as those in the cross-planes method. In addition, the threshold in the quadri-planes method for dichotomization of the final CAD assessment was set at any 2 of 4 positive planes; however, the threshold in the cross-planes method was set as any 1 of 2 positive planes. This may have led to the increase in specificity and the decrease in sensitivity of the quadri-planes method compared to the cross-planes method.

When the assessments differed in category 4A and 4B between the readers and the CAD, the proportion of correct adjustments using CAD for the inexperienced reader was higher than that for the experienced reader (81/115, 70% vs 37/70, 53%). This indicates that the less experienced reader obtained more benefit from the CAD system; this is related to the fact that combining CAD with the quadri-planes method led to improvements in sensitivity, specificity, and accuracy for the inexperienced reader, while the accuracy between the quadri-planes method and the experienced reader was similar (83.8% vs 86.3%). For CAD with the quadri-planes method, the consistency between the experienced reader and novice reader was good. As such, CAD assistance with the quadri-planes method can not only improve diagnostic performance but can also be expected to play a more weighted role in providing a second opinion, especially for less experienced readers. Consequently, this system can reduce misdiagnosis by less experienced readers in addition to reducing variability in readers' interpretations and overcoming the effects of inexperience. These improvements in diagnostic performance by combining CAD and ultrasound may reduce both the misdiagnosis and missed diagnosis ratios of breast cancer by readers with different experience levels.

Several reports have described applying different types of CAD to breast ultrasound [6,19-21]. These studies all reported that the CAD systems enhanced the diagnostic performance of breast ultrasound, especially specificity and accuracy. Shen et al [20] argued that CAD systems can be helpful in evaluating fuzzy category 4 lesions. Wang et al [21] suggested that combining CAD is more helpful for inexperienced readers than experienced ones, with greater improvement in the diagnostic performance in the inexperienced group. Kim's study involved 2 staff

radiologists with 7 and 19 years of experience, respectively; both these readers can be described as experienced, so their false positive rates were low and their false negative rates were relatively high. In addition, the surgery proportion was only 27.6% and the core needle biopsy proportion was 61.5%, which may also have affected the results. The retrospective analysis was performed by only 1 radiologist with 7 years of ultrasound experience. In Wang's study, the CAD system was relatively old, and the experience between the 8 readers varied, which may have led to increases in false negative and false positive rates. Therefore, the results of the above studies show that traditional CAD methods are not sufficient to balance the sensitivity and specificity to effectively reduce false negative or false positive results. In our study, the sensitivity, NPV, and accuracy of both readers improved; this supports the idea that S-Detect can reliably provide a second view that can be referred to by readers. Although the CAD methods were not exactly the same as in previous studies [17,18], high sensitivity balanced with specificity is a remarkable superiority of the quadri-planes method. Instead, the proportion of the benign lesions in our study was lower (Table 1), and the mean size of the lesions was larger; also, all the patients had breast masses as their prominent symptom, which may lead to differences between the results as in the study by Wu et al [22]. According to the BI-RADS criteria [23], we subdivided category 4 into categories 4A, 4B, and 4C, and the threshold was set to category 4A in grayscale ultrasound; consequently, the specificity was high and the sensitivity was relatively low. Moreover, S-Detect provides the final assessment in a dichotomized form of possibly benign and possibly malignant; we consider that these factors may also affect the diagnostic capability of readers combined with CAD.

Strengths and Prospects

The results of our study are encouraging for daily clinical breast cancer screening practiced by readers, although some pathology subtypes of breast cancer had better outcomes in situ [24]. However, breast cancer is still a relatively aggressive disease that possesses higher rates of metastases and poorer survival rates [25]. Thus, it is important to detect breast cancer accurately in early stages to reduce its mortality rate [26]. Additionally, S-Detect is a concise and user-friendly program that is integrated in the ultrasound machine; the quadri-planes method enables the reader to immediately achieve a more precise result during real-time ultrasonography, which can easily be applied to routine work (Figure 3). However, it is not recommended to apply CAD alone or as a replacement for a human reader in the diagnosis of breast lesions at present, as shown in Kim's study [6] (Figure 4). However, there is reason to believe that this will be possible in the near future. Further investigation with technical advances can be anticipated to develop a more sophisticated algorithm using the multiple plane assessment BI-RADS ultrasonographic categories.

Ultrasound scanning is a real-time and multi-angle inspection process; a lesion can be observed from different planes to collect imaging features such as the internal situation, the relationship of the lesion with its surroundings, the blood supply model, and patient histories. Obviously, ultrasound can obtain more image data and clinical information than CAD. The quadri-planes method with CAD can extract more features from a tumor with maximum objectivity; combined with the expertise of a reader, the weaknesses of each method can be counteracted by the strengths of the others, which can assist readers in making more accurate diagnoses regardless of their experience.

Figure 3. A breast lesion assessed by CAD with the cross-planes method (A, B) and the quadri-planes method (A-D).

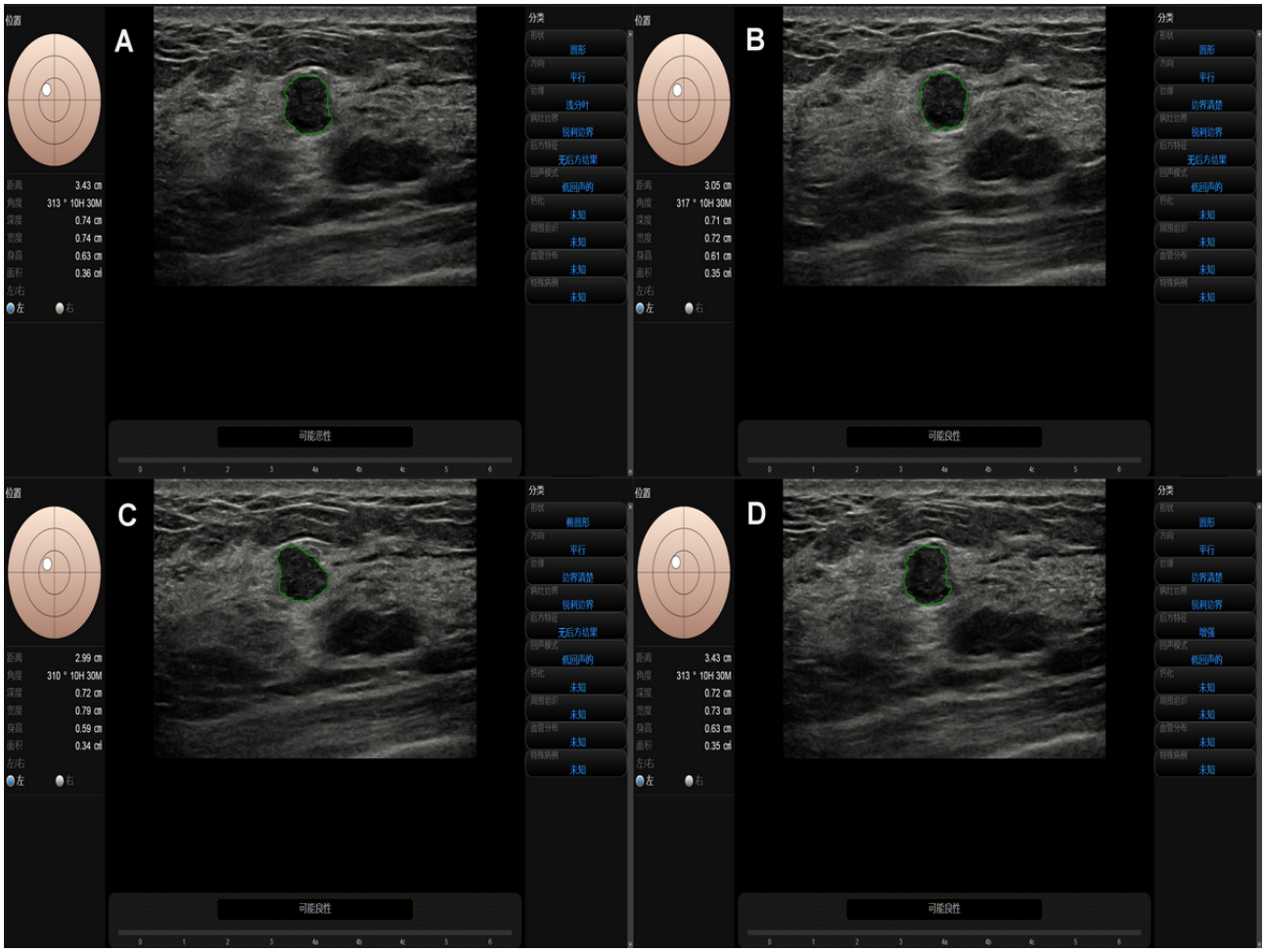
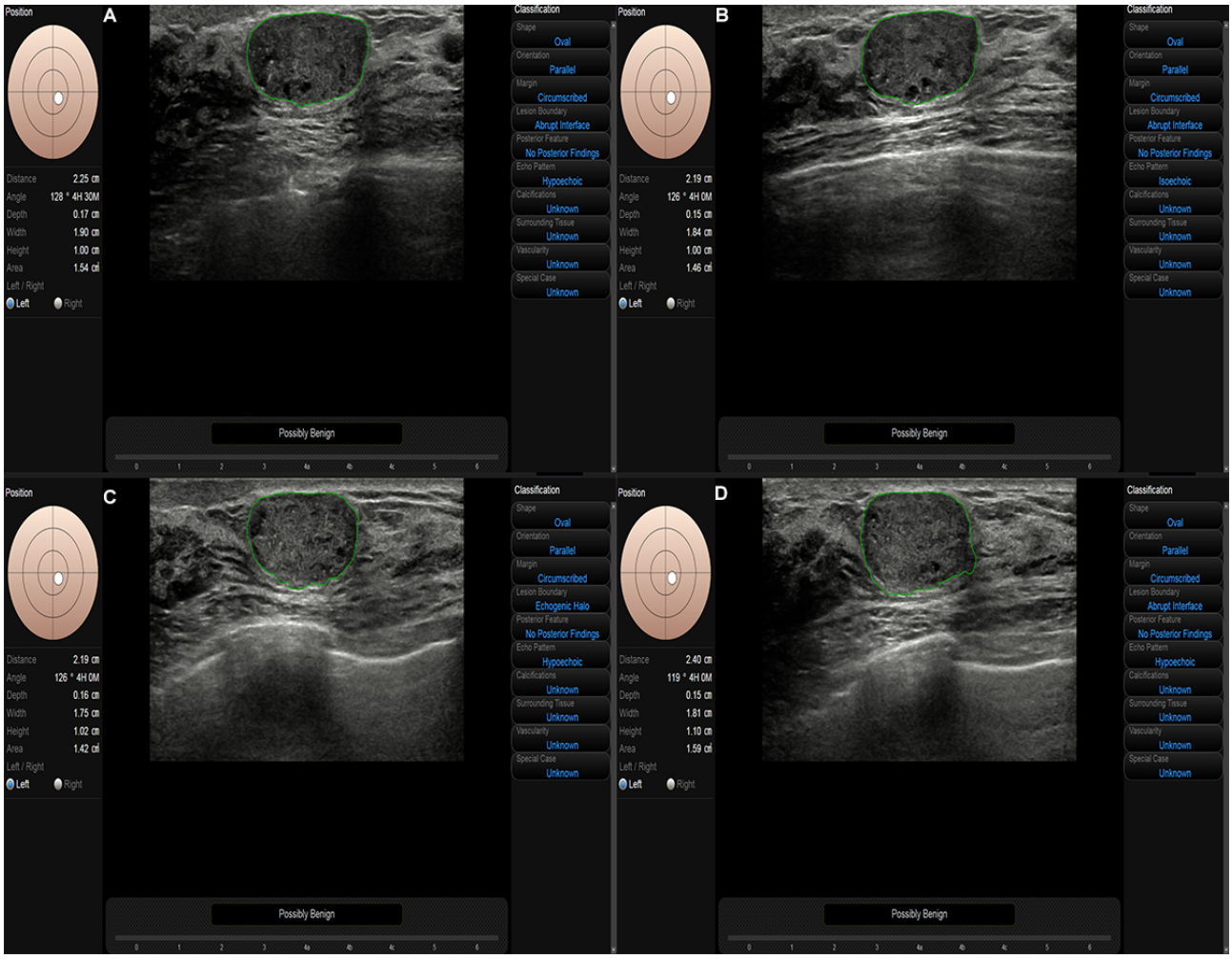


Figure 4. Example of a ductal carcinoma in situ lesion with a size of 1.90×1.10 centimeters showing a clear margin, regular shape, and microcalcification that was incorrectly diagnosed as benign by S-Detect with the cross-planes (A, B) and quadri-planes (A-D) methods. The lesion was classified as BI-RADS category 4B by the readers.



Limitations

There are several limitations of this study. First, the number of cases in the single center study was relatively small. Second, the presentation of calcifications is still not included in the current version of S-Detect due to its limited analysis ability for microcalcifications [27]. Third, some small nodules classified as BI-RADS category 2 or 3 with sizes of around 1 cm without surgical operation were not included in this study, which may have affected the results. Fourth, the number of planes of the lesions for CAD was set to 4. It can be argued that it would have been better to study additional planes. Fifth, both of the readers were relatively inexperienced breast scan readers. In China, the specialty of breast imaging is new, and its staff are young compared with those in other imaging specialties. These factors may have affected the results.

Conclusion

S-Detect is a feasible diagnostic tool that can improve the sensitivity, accuracy, and AUC of both novice and experienced readers in the quadri-planes method while also improving the specificity for the novice reader; thus, it demonstrates important application value in the clinical diagnosis of breast cancer.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (81871367).

Conflicts of Interest

None declared.

References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019 Jan;69(1):7-34 [FREE Full text] [doi: 10.3322/caac.21551] [Medline: 30620402]

2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66(2):115-132 [FREE Full text] [doi: 10.3322/caac.21338] [Medline: 26808342]
3. Brinkley D, Haybrittle JL. The curability of breast cancer. Lancet 1975 Jul 19;2(7925):95-97 [FREE Full text] [doi: 10.1016/s0140-6736(75)90003-3] [Medline: 49738]
4. Takahashi R, Kajikawa Y. Computer-aided diagnosis: A survey with bibliometric analysis. Int J Med Inform 2017 May;101:58-67 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.02.004] [Medline: 28347448]
5. Cho E, Kim E, Song MK, Yoon JH. Application of Computer-Aided Diagnosis on Breast Ultrasonography: Evaluation of Diagnostic Performances and Agreement of Radiologists According to Different Levels of Experience. J Ultrasound Med 2018 Jan;37(1):209-216 [FREE Full text] [doi: 10.1002/jum.14332] [Medline: 28762552]
6. Kim K, Song MK, Kim E, Yoon JH. Clinical application of S-Detect to breast masses on ultrasonography: a study evaluating the diagnostic performance and agreement with a dedicated breast radiologist. Ultrasonography 2017 Jan;36(1):3-9 [FREE Full text] [doi: 10.14366/usg.16012] [Medline: 27184656]
7. Horsch K, Giger ML, Vyborny CJ, Venta LA. Performance of computer-aided diagnosis in the interpretation of lesions on breast sonography. Acad Radiol 2004 Mar;11(3):272-280. [doi: 10.1016/s1076-6332(03)00719-0] [Medline: 15035517]
8. Komeda Y, Handa H, Watanabe T, Nomura T, Kitahashi M, Sakurai T, et al. Computer-Aided Diagnosis Based on Convolutional Neural Network System for Colorectal Polyp Classification: Preliminary Experience. Oncology 2017;93 Suppl 1:30-34 [FREE Full text] [doi: 10.1159/000481227] [Medline: 29258081]
9. Zhao C, Xiao M, Jiang Y, Liu H, Wang M, Wang H, et al. Feasibility of computer-assisted diagnosis for breast ultrasound: the results of the diagnostic performance of S-detect from a single center in China. Cancer Manag Res 2019;11:921-930 [FREE Full text] [doi: 10.2147/CMAR.S190966] [Medline: 30774422]
10. Choi J, Kang BJ, Baek JE, Lee HS, Kim SH. Application of computer-aided diagnosis in breast ultrasound interpretation: improvements in diagnostic performance according to reader experience. Ultrasonography 2018 Jul;37(3):217-225 [FREE Full text] [doi: 10.14366/usg.17046] [Medline: 28992680]
11. Migowski A. [Early detection of breast cancer and the interpretation of results of survival studies]. Cien Saude Colet 2015 Apr;20(4):1309 [FREE Full text] [doi: 10.1590/1413-81232015204.17772014] [Medline: 25923642]
12. Guo R, Lu G, Qin B, Fei B. Ultrasound Imaging Technologies for Breast Cancer Detection and Management: A Review. Ultrasound Med Biol 2018 Jan;44(1):37-70 [FREE Full text] [doi: 10.1016/j.ultrasmedbio.2017.09.012] [Medline: 29107353]
13. Wang M, Yang Z, Liu C, Yan J, Zhang W, Sun J, et al. Differential Diagnosis of Breast Category 3 and 4 Nodules Through BI-RADS Classification in Conjunction with Shear Wave Elastography. Ultrasound Med Biol 2017 Mar;43(3):601-606 [FREE Full text] [doi: 10.1016/j.ultrasmedbio.2016.10.004] [Medline: 27988221]
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]
15. Waks AG, Winer EP. Breast Cancer Treatment. JAMA 2019 Jan 22;321(3):316. [doi: 10.1001/jama.2018.20751] [Medline: 30667503]
16. Yongping L, Zhou P, Juan Z, Yongfeng Z, Liu W, Shi Y. Performance of Computer-Aided Diagnosis in Ultrasonography for Detection of Breast Lesions Less and More Than 2 cm: Prospective Comparative Study. JMIR Med Inform 2020 Mar 02;8(3):e16334 [FREE Full text] [doi: 10.2196/16334] [Medline: 32130149]
17. Chabi M, Borget I, Ardiles R, Aboud G, Boussouar S, Vilar V, et al. Evaluation of the accuracy of a computer-aided diagnosis (CAD) system in breast ultrasound according to the radiologist's experience. Acad Radiol 2012 Mar;19(3):311-319. [doi: 10.1016/j.acra.2011.10.023] [Medline: 22310523]
18. Morra L, Sacchetto D, Durando M, Agliozzo S, Carbonaro LA, Delsanto S, et al. Breast Cancer: Computer-aided Detection with Digital Breast Tomosynthesis. Radiology 2015 Oct;277(1):56-63. [doi: 10.1148/radiol.2015141959] [Medline: 25961633]
19. Xiao M, Zhao C, Zhu Q, Zhang J, Liu H, Li J, et al. An investigation of the classification accuracy of a deep learning framework-based computer-aided diagnosis system in different pathological types of breast lesions. J Thorac Dis 2019 Dec;11(12):5023-5031 [FREE Full text] [doi: 10.21037/jtd.2019.12.10] [Medline: 32030218]
20. Shen W, Chang R, Moon WK. Computer aided classification system for breast ultrasound based on Breast Imaging Reporting and Data System (BI-RADS). Ultrasound Med Biol 2007 Nov;33(11):1688-1698. [doi: 10.1016/j.ultrasmedbio.2007.05.016] [Medline: 17681678]
21. Wang Y, Jiang S, Wang H, Guo YH, Liu B, Hou Y, et al. CAD algorithms for solid breast masses discrimination: evaluation of the accuracy and interobserver variability. Ultrasound Med Biol 2010 Aug;36(8):1273-1281. [doi: 10.1016/j.ultrasmedbio.2010.05.010] [Medline: 20691917]
22. Wu J, Zhao Z, Zhang W, Liang M, Ou B, Yang H, et al. Computer-Aided Diagnosis of Solid Breast Lesions With Ultrasound: Factors Associated With False-negative and False-positive Results. J Ultrasound Med 2019 Dec;38(12):3193-3202 [FREE Full text] [doi: 10.1002/jum.15020] [Medline: 31077414]
23. Rao AA, Feneis J, Lalonde C, Ojeda-Fournier H. A Pictorial Review of Changes in the BI-RADS Fifth Edition. Radiographics 2016;36(3):623-639 [FREE Full text] [doi: 10.1148/rg.2016150178] [Medline: 27082663]
24. Ward EM, DeSantis CE, Lin CC, Kramer JL, Jemal A, Kohler B, et al. Cancer statistics: Breast cancer in situ. CA Cancer J Clin 2015;65(6):481-495 [FREE Full text] [doi: 10.3322/caac.21321] [Medline: 26431342]

25. Harbeck N, Gnant M. Breast cancer. Lancet 2017 Mar 18;389(10074):1134-1150. [doi: 10.1016/S0140-6736(16)31891-8] [Medline: 27865536]
26. Wang Y, Fan W, Zhao S, Zhang K, Zhang L, Zhang P, et al. Qualitative, quantitative and combination score systems in differential diagnosis of breast lesions by contrast-enhanced ultrasound. Eur J Radiol 2016 Jan;85(1):48-54. [doi: 10.1016/j.ejrad.2015.10.017] [Medline: 26724648]
27. Sickles EA. Breast calcifications: mammographic evaluation. Radiology 1986 Aug;160(2):289-293. [doi: 10.1148/radiology.160.2.3726103] [Medline: 3726103]

Abbreviations

AUC: area under the receiving operating characteristic curve
BI-RADS: Breast Imaging Reporting and Data System
CAD: computer-aided diagnosis
NPV: negative predictive value
PPV: positive predictive value
ROC: receiver operating characteristic

Edited by G Eysenbach; submitted 14.02.20; peer-reviewed by J Zhang, K Karanam, D Di Stasio, T Muto; comments to author 10.03.20; revised version received 16.03.20; accepted 10.04.20; published 05.05.20.

Please cite as:
Yongping L, Juan Z, Zhou P, Yongfeng Z, Liu W, Shi Y
Evaluation of the Quadri-Planes Method in Computer-Aided Diagnosis of Breast Lesions by Ultrasonography: Prospective Single-Center Study
JMIR Med Inform 2020;8(5):e18251
URL: https://medinform.jmir.org/2020/5/e18251
doi:10.2196/18251
PMID:32369039

Original Paper

Use of Machine Learning Techniques for Case-Detection of Varicella Zoster Using Routinely Collected Textual Ambulatory Records: Pilot Observational Study

Corrado Lanera[1], MSc, PhD; Paola Berchialla[2], MSc, PhD; Ileana Baldi[1], MSc, PhD; Giulia Lorenzoni[1], MA, PhD; Lara Tramontan[3], PhD; Antonio Scamarcia[4], MD; Luigi Cantarutti[4], MD; Carlo Giaquinto[5], PhD; Dario Gregori[1], MA, PhD

[1]Department of Cardiac Thoracic Vascular Sciences and Public Health, University of Padova, Unit of Biostatistics, Epidemiology and Public Health, Padova, Italy

[2]Department of Clinical and Biological Science, University of Turin, Torino, Italy

[3]Arsenàl.IT, Treviso, Italy

[4]Società Servizi Telematici, Pedianet, Padova, Italy

[5]Department of Women's and Children's Health, University of Padova, Padova, Italy

Corresponding Author:
Dario Gregori, MA, PhD
Department of Cardiac Thoracic Vascular Sciences and Public Health
University of Padova
Unit of Biostatistics, Epidemiology and Public Health
Via Leonardo Loredan 18
Padova, 35121
Italy
Phone: 39 049 827 5384
Fax: 39 049 827 5407
Email: dario.gregori@unipd.it

Abstract

Background: The detection of infectious diseases through the analysis of free text on electronic health reports (EHRs) can provide prompt and accurate background information for the implementation of preventative measures, such as advertising and monitoring the effectiveness of vaccination campaigns.

Objective: The purpose of this paper is to compare machine learning techniques in their application to EHR analysis for disease detection.

Methods: The Pedianet database was used as a data source for a real-world scenario on the identification of cases of varicella. The models' training and test sets were based on two different Italian regions' (Veneto and Sicilia) data sets of 7631 patients and 1,230,355 records, and 2347 patients and 569,926 records, respectively, for whom a gold standard of varicella diagnosis was available. Elastic-net regularized generalized linear model (GLMNet), maximum entropy (MAXENT), and LogitBoost (boosting) algorithms were implemented in a supervised environment and 5-fold cross-validated. The document-term matrix generated by the training set involves a dictionary of 1,871,532 tokens. The analysis was conducted on a subset of 29,096 tokens, corresponding to a matrix with no more than a 99% sparsity ratio.

Results: The highest predictive values were achieved through boosting (positive predicative value [PPV] 63.1, 95% CI 42.7-83.5 and negative predicative value [NPV] 98.8, 95% CI 98.3-99.3). GLMNet delivered superior predictive capability compared to MAXENT (PPV 24.5% and NPV 98.3% vs PPV 11.0% and NPV 98.0%). MAXENT and GLMNet predictions weakly agree with each other (agreement coefficient 1 [AC1]=0.60, 95% CI 0.58-0.62), as well as with LogitBoost (MAXENT: AC1=0.64, 95% CI 0.63-0.66 and GLMNet: AC1=0.53, 95% CI 0.51-0.55).

Conclusions: Boosting has demonstrated promising performance in large-scale EHR-based infectious disease identification.

(JMIR Med Inform 2020;8(5):e14330) doi:10.2196/14330

KEYWORDS

machine learning technique; text mining; electronic health report; varicella zoster; pediatric infectious disease

Introduction

Improving the predictive capability of infectious disease detection at the population level is an important public health issue that can provide the background information necessary for the implementation of effective control strategies, such as advertising and monitoring the effectiveness of vaccination campaigns [1].

The need for fast, cost-effective, and accurate detection of infection rates has been widely investigated in recent literature [2]. Particularly, the combination of increased electronic health report (EHR) implementation in primary care, the growing availability of digital information within the EHR, and the development of data mining techniques offer great promise for accelerating pediatric infectious disease research [3].

Although EHR data are collected prospectively in real time at the point of health care delivery, observational studies intended to retrospectively assess the impact of clinical decisions are likely the most common type of EHR-enabled research [3].

Among the high-impact diseases, the prompt identification of varicella zoster viral infections is of key interest due to the debate around the need and cost-benefit dynamics of a mass-vaccination program for young children [4,5].

Challenges in this context arise from both the unique epidemiological characteristics of varicella zoster with respect to information extraction, such as age-specific consultation rates, seasonality, force of infection, hospitalization rates, and inpatient days [6], and from the way that medical records are organized, often in free-format and uncoded fields [7]. A critical step is to transform this large amount of health care data into knowledge.

Data extraction from free text for disease detection at the individual level can be based on manual, in-depth examinations of individual medical records or, to contain costs and ensure time-tightening and control, by automatic coding. Machine learning techniques (MLTs) are the most commonly used approaches [8] and show good overall performance [9,10]. Nevertheless, few indications are currently available on the most appropriate technique to use, and comparative evidence is still lacking on the performances of each available technique [11] in the field of pediatric infectious disease research.

In recent years, generalized linear model (GLM)-based techniques have been largely used for the text mining of EHRs, both as a technique of choice [12] and as a benchmark [13]. The performance of GLMs, especially multinomial or in the simplest cases logistic regression, has been indicated as unsatisfactory [14] because they are prone to overfitting and are sensitive to outliers. Enhancements to GLMs have been proposed recently in the form of the lasso and elastic-net regularized GLM [15] (GLMNet), multinomial logistic regression (maximum entropy [MAXENT]), and the boosting approach implemented in the LogitBoost algorithm [16] to overcome the limitations of naïve GLMs. Nevertheless, to the best of our knowledge, no comparisons have been made among these techniques to determine to what extent improvements are needed.

The purpose of this study is to make comparisons among enhanced GLM techniques in the setting of automatic disease detection [17]. Particularly, these methods will be assessed on their ability of identifying cases of varicella from a large set of EHRs.

Methods

Electronic Medical Record Database

The Italian Pedianet database [18] collects anonymized clinical data from more than 300 pediatricians throughout the country. This database focuses on children 0-14 years of age [19-22] and records the reasons for accessing health care, diagnosis, and clinical details. The sources of those data are primary care records written in Italian, which are filled in by pediatricians with clinical details about diagnosis and prescriptions; they also contain details about the eventual hospitalization and specialist referrals.

For the purpose of this study, we were allowed to access only two subsets of the Pedianet database, corresponding to the data collected between 2004 and 2014 in the Italian regions of Veneto (northern Italy) and Sicilia (South Italy). Since the Veneto region data set was larger, it was considered for carrying out the training of the model. The data set of the Sicilia region provided an independent data set for testing the model. The main characteristics of the two data sets are reported in Table 1. It is worth noting that the proportion of positive cases of varicella is different in the two databases. Interpreting differences in prevalence between regions is beyond the purpose of this study; nevertheless, given the smaller prevalence, there is an expected lower positive predictive value (PPV) and a higher negative predictive value (NPV) on the test set.

The Pedianet source data includes five different tables. In Table 2, we report a short description of them.

Table 1. Main characteristics used for the train (Veneto) and test (Sicilia) data sets.

Characteristic	Train	Test
Database	Pedianet	Pedianet
Language	Italian	Italian
Italian Region	Veneto	Sicilia
Date span	January 2, 2004-December 31, 2014	January 7, 2004-December 30, 2014
Records, n	1,230,355	569,926
Children, n	7631	2347
Pediatricians, n	46	13
Positive cases, n (%)	3481 (45.6%)	128 (5.4%)

Table 2. Tables used from the Pedianet database.

Table topic	Content	Type of data	Example
Accessing	Reasons for accessing the pediatrician and diagnoses	Free text (including codes)	• Ritardo di crescita <783.4>
Diaries	Pediatrician's free-text diaries	Free text	• DIBASE OS GTT 10ML 10000UI/ML n° conf. 2\r\n per Visita di controllo e di follow up\r\n\r\n
Hospitalizations	Details on hospital admissions, diagnoses, and length of stays	Free text	• Divisione di pediatria • Tosse, difficolta' respiratoria e di alimentazione
SOAP[a]	Symptoms, objectivity, diagnosis, or prescriptions	Free text (including codes)	• *SOAP*[b]: "P", • *SOAP_code*: "77469", • *SOAP_text*: "visita otorinolaringoiatrica<89.7>"
Specialistic visits	Visit type and its diagnosis	Free text including (codes)	• *codice_visitaSP*: "89.01", • *visita*: "ecografia anche sec. Graaf per screening", • *diagnosi*: "problemi della vista <V41.0>"

[a]SOAP: symptoms, objectivity, diagnosis, or prescriptions.

[b]For tables with multiple fields, field names are reported in italics.

All the tables can be linked at the individual level (ie, each row of all the tables contains the fields for reporting information on dates, the assisting pediatrician's anonymous identifier, and the patients' anonymous identifier, which constitutes the linking key).

Case Definition

The case definition comes directly from the gold standard provided, and the training set for machine learning was created using those dichotomous labels (ie, 0=noncase, that is not a varicella case; and 1=case, that is a varicella case).

Training and Test Sets for Machine Learning

Linking by patient ID, pediatrician ID, and reporting date, we merged the five tables into a single table consisting of several entries, each of which represents a visit or evaluation of a patient carried out by a pediatrician on a specific day. At this step, the information (excluding patient ID, pediatrician ID, and reporting date) is contained in 15 columns containing free text mixed with coded text, which was considered by us as free text as well. Finally, all remaining columns of the table were merged into a single corpus (ie, a body of text). This process was applied to train the models on 1,230,355 entries (database of the Veneto region) and to test them on 569,926 entries (database of the Sicily region) separately.

Preprocessing

Text analysis by a computer program is possible only after establishing a way to convert text (ie, readable to humans) into numbers (ie, readable to computers). This process is called preprocessing, and it is the first [23] and probably the most important step in data mining [24]. To process the corpus of Pedianet EHRs included in the training set, we used the following strategy. First, we converted all fields in a text type; lowered the content; and cleared it of symbols, punctuation, numbers, and extra white spaces. Second, we stemmed the words (ie, reducing them to their basic form, or "root"), which is recognized as one of the most important procedures to perform [25], and constructed 2-gram tokens, which has been shown to be the optimal rank for gram tokenization [26]. Third, we removed all the (stemmed) stop words (ie, common and nonmeaningful words such as articles or conjunctions) from the set of tokens as well as all bigrams containing any of them. We chose this strategy after exploring different approaches described in [27]. Fourth, we created the document-term matrix (DTM) as a patient-token matrix. To consider both the importance of the tokens within a patient (ie, one row of the DTM) and its

discrimination power between patients' records (ie, the rows of the DTM), we computed the TF-iDF (term frequencies–inverse document frequencies) weights. TF-iDF weights help to adjust for the presence of words that are more frequent but less meaningful [28]. TF-iDF-ij entry is equal to the product of the frequency of the j-th token in the i-th document by the logarithm of the inverse of the number of documents that contain that token (ie, the more frequent a word appears in a document the more its weight rises for that document), and the more documents that contain the j-th token, the more the weight shrinks across all the documents [29]. In the initial DTM there were 1,871,532 tokens that appear at least once, with a nonsparse/sparse entries ratio of (18,951,304/14,262,709,388). We decided to reduce it to achieve a maximum of 99% overall sparsity. Filtering out the tokens that do not appear in at least 1% of the documents had reduced it down to 94% (ie, 29,096 tokens that appear at least once for a nonsparse/sparse entries ratio of 13,140,370/208,891,206). The choice of a 99% level of sparsity was a tradeoff between the need to retain as many tokens as possible and the computational effort.

The corpus of Pedianet EHRs comprised in the test set went through the same text preprocessing strategy in the same order, and then the DTM was created with the initial TF weighing scheme. Furthermore, it was adapted with the same tokens retained in the training phase (ie, adding the missing tokens, weighting them as zero, and removing the ones not included in the training DTM) and was finally reweighted with the TF-iDF weighing scheme with the same retained iDF weights of the corresponding training DTM, which were retained when applied to the whole training data set. Those are necessary steps to guarantee that the two feature spaces are the same and that the models trained can be evaluated on the test set.

Machine Learning Techniques

Enhancements of GLMs for carrying out text mining on EHRs have been proposed in the form of the lasso and GLMNet [16], multinomial logistic regression (MAXENT), and the boosting approach (LogitBoost) [16].

GLMNet is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods applied in synergy with a link function and a variance function to overcome linear model limitations (eg, the constant variability among the mean and the normality of the data). The link function selected was the binomial (ie, the model fit a regularized logistic regression model for the log odds), while the amount of regularization was automatically selected by the algorithm through an exploration of 100 values between the minimum value that reduced all the coefficients to zero and its 0.01 fraction.

MAXENT is an implementation of (multinomial) logistic regression aimed at minimizing the memory load on large data sets in R (R Foundation for Statistical Computing) and is primarily designed to work with the sparse DTM provided by the R package [30]. It has been proven to provide results mathematically equivalent to a GLM with a Poisson link function [31].

Boosting is a general approach for improving the predictive capability of any given learning algorithm. We used the adaptations of Tuszynski [32] to the original algorithm, (ie, LogitBoost [33,34]), which is aimed at making the entire process more efficient while applying it on large data sets. The standard boosting technique [34] is applied to the sequential use of a decision stump classification algorithm as a weak learner (ie, a single binary decision tree). The number of stumps considered is the same as the columns provided in the training set.

Those techniques are chosen among computationally treatable algorithms for use with large data sets [30]. GLMNet and MAXENT represent classical benchmark approaches to linear and logistic classification, respectively, in a manner that differs from LogitBoost, which is a modern boosted tree-based machine learning approach [35,36]. Moreover, LogitBoost generalizes the classical logistic models by fitting a logistic model at each node [37] and shows an alternative point of view with regards to models such as the GLMs, for which the structure of the learner must be chosen a priori [38].

Training and Testing

We addressed the issue of internal validation by performing cross-validation on the training set comprising records from the Veneto region. We dealt with external validation by accessing a truly external sample of Pedianet EHRs from another Italian region, Sicily. This accomplishes two tasks: preserving precision in the training phase and complementing study findings with external validation results using data that were not available when the predictive tool was developed.

We used a 5-fold cross-validation approach to validate each of the three MLTs on the DTM with the corresponding (by row) "case/non-case" attached labels. All MLTs were simultaneously fitted on the same set of folds to ensure a proper comparison between techniques. Values of k=10 or k=5 (especially for large data sets) have been shown empirically to yield acceptable (in terms of bias-variance trade-off) error rates [39,40]. Thus, the choice of 5-folds was driven by the computational complexity, the fewer folds, the less complexity.

As measures of performance, we calculated point estimates and 95% CIs for the following.

- PPV or Precision:  , that is the fraction of positively identified cases that are true positives
- NPV:  , that is the fraction of positively identified noncases that are true negatives
- Sensitivity or Recall:  , that is the true positive rate
- Specificity:  , that is the true negative rate
- F score:  , the harmonic mean of the PPV (Precision) and Sensitivity (Recall)

The Gwet agreement coefficient 1 (AC1) statistics of agreement [41,42] between the techniques were computed and reported, along with their corresponding 95% CIs. Given that A=the number of times both models classify a record as noncase, D=the number of times both models classify a record as a case, and N=the total sample size, then  , where , and e^{γ} is the

agreement probability by chance and is equal to $2q(1-q)$, where  ; A1 is the number of records classified as noncase by model 1, and B1 is the number of records classified as noncase by model 2. AC1 has been used given its propensity to be weakly affected by marginal probability, and therefore it was chosen to manage unbalanced data [43].

All the analyses were implemented in the R system [44] with the computing facilities of the Unit of Biostatistics, Epidemiology and Public Health. The R packages used were: *SnowballC* (to stem the words) and *RWeka* (to create n-grams) for the preprocessing step; *Matrix* and *SparseM* to manage sparse matrices; *GLMNet*, *MAXENT*, and *caTools* for the GLMNet, MAXENT, and LogitBoost MLT implementation; *caret* to create and evaluate the cross-validation folds; *ROCR* to estimate the performance; and the *tidyverse* bundle of packages for data management, functional programming, and plots. A git repository of the analysis code is available [45].

Results

The flow chart, from data acquisition to preprocessing, is shown in Figure 1. In the training set, 29,096 initial terms out of 1,871,532 were retained by the sparsity reduction step. Boosting significantly outperforms all other MLTs on the training set, with the highest *F* score and PPV. The GLMNet predictor delivered a superior *F* score and greater PPV compared to MAXENT (Table 3). The same results held on the test set (Table 4) and agreement between MLT predictions on the training set was good as measured by AC1 statistics (Table 5).

Figure 1. Flowchart from the acquisition of the five tables containing the electronic health records (dark gray) in the training set that were merged into a single table (dark blue); preprocessed (gray) with the specification of what was removed (pink) prior to the creation of the document-term matrix (DTM) (yellow); the computation of the weights (light blue); the dimensionality reduction, that is the reduction of the terms used (light gray), and the final DTM used (green). DTM: document-term matrix; SOAP: symptoms, objectivity, diagnosis, or prescriptions; TF-iDF: term frequencies–inverse document frequencies.

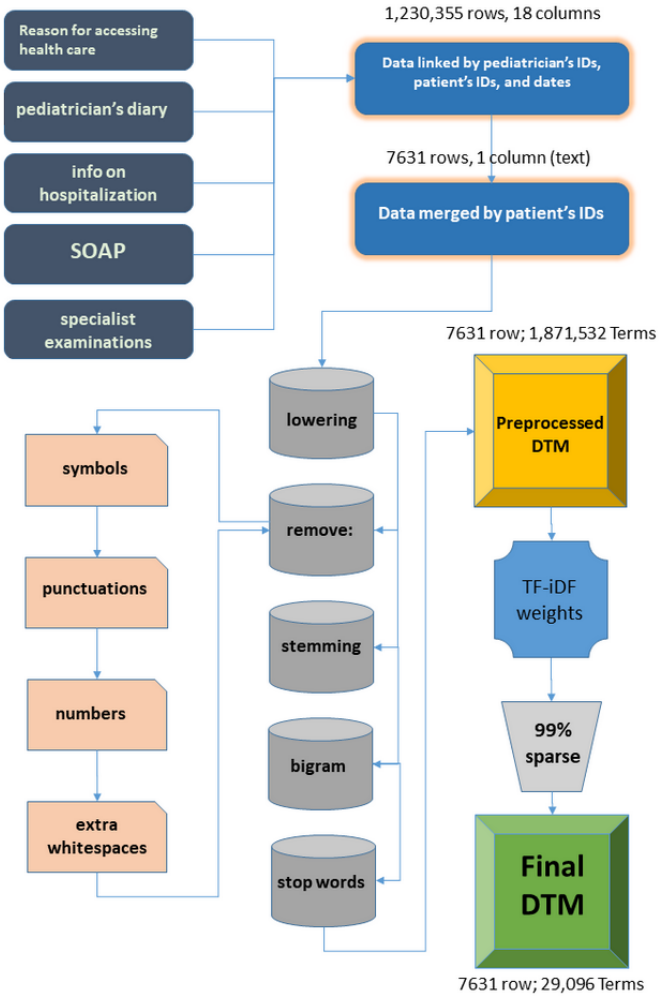


Table 3. Performance on the training set of the three machine learning techniques using a 5-fold cross-validation method.

Technique	Sensitivity, mean (95% CI)	PPV[a], mean (95% CI)	NPV[b], mean (95% CI)	Specificity, mean (95% CI)	*F* score, mean (95% CI)
GLMNet[c]	80.2 (77.7-82.7)	73.2 (70.9-75.6)	90.9 (89.6-92.2)	87.1 (85.6-88.7)	76.5 (75.6-77.5)
MAXENT[d]	68.8 (66.8-70.7)	66.0 (62.5-69.5)	86.1 (85.2-86.9)	84.5 (82.7-86.3)	67.4 (64.7-70.0)
Boosting	86.6 (82.1-91.1)	95.8 (93.2-98.5)	94.4 (92.4-96.3)	98.3 (97.0-99.6)	90.9 (89.7-92.1)

[a]PPV: positive predicative value.

[b]NPV: negative predicative value.

[c]GLMNet: elastic-net regularized generalized linear model.

[d]MAXENT: maximum entropy.

Table 4. Performance on the test set of the three machine learning techniques under consideration.

Technique	Sensitivity, mean (95% CI)	PPV[a], mean (95% CI)	NPV[b], mean (95% CI)	Specificity, mean (95% CI)	*F* score, mean (95% CI)
GLMNet[c]	72.3 (66.4-78.1)	24.5 (21.0-28.0)	98.3 (97.9-98.6)	87.4 (85.4-89.5)	36.5 (32.2-40.8)
MAXENT[d]	74.8 (62.2-87.5)	11.0 (9.5-12.5)	98.0 (97.3-98.6)	65.5 (54.7-76.2)	19.1 (17.2-20.9)
Boosting	79.2 (69.7-88.7)	63.1 (42.7-83.5)	98.8 (98.3-99.3)	96.9 (94.2-99.6)	68.5 (59.3-77.7)

[a]PPV: positive predicative value.

[b]NPV: negative predicative value.

[c]GLMNet: elastic-net regularized generalized linear model.

[d]MAXENT: maximum entropy.

Table 5. Agreement between elastic-net regularized generalized linear model, maximum entropy, and boosting using 5-fold cross-validation.

Technique	Wrongly agree[a], n	Correctly agree[b], n	Disagree[c], n	Gwet AC1[d,e] (95% CI)
GLMNet[f] vs MAXENT[g]	669	5609	1353	0.68 (0.67-0.70)
GLMNet vs boosting	195	6269	1146	0.74 (0.72-0.75)
MAXENT vs boosting	224	5895	1491	0.66 (0.65-0.68)

[a]The "Wrongly Agree" column refers to the number of records misclassified by both techniques.

[b]The "Correctly Agree" column states the number of records correctly classified by both techniques.

[c]The "Disagree" column lists the number of records for which the techniques disagree in the classification.

[d]AC1: agreement coefficient 1.

[e]Gwet AC1 represents the index of agreement between the identified techniques. Legend for AC1 is: AC1<0=disagreement; AC1 0.00-0.40=poor; AC1 0.41-0.60=discrete; AC1 0.61-0.80=good; AC1 0.81-1.00=optimal.

[f]GLMNet: elastic-net regularized generalized linear model.

[g]MAXENT: maximum entropy.

With the aim to analyze the most relevant errors, we explored if any records were wrongly classified by all the techniques. There were 3 records: 1 wrongly classified as positive and 2 wrongly classified as negatives by all the MLTs.

Discussion

Principal Findings

The application of MLTs to EHRs constitutes the analytical component of an emerging research paradigm that rests on the capture and preprocessing of massive amounts of clinical data to gain clinical insights and ideally to complement the decision-making process at different levels, from individual treatment to definition of national public health policies. As acknowledged by others [46], the development and application of big data analysis methods on EHRs may help create a continually learning health care system [47].

This study trains and compares three different machine learning approaches towards infectious disease detection at the population level based on clinical data collected in primary care EHRs. In line with the recommended paradigm for model validation [39], the MLTs' performance underwent internal validation through cross-validation and external validation on an independent set of EHRs.

The predictive capabilities of the developed MLTs are promising even if quite different from each other (eg, validation *F* scores range from 67%-91% and test *F* scores range from 19%-69%). Findings on the better performance reached by LogitBoost are in line with recent evidence that shows an improvement in general classification problems moving from MAXENT algorithms to LogitBoost-based ones [48]. LogitBoost is thus

confirmed to be a useful technique for solving health-related classification problems [34].

Only three records were wrongly classified by all the models. The first one was wrongly classified as positive probably because the text entry was "vaccini:varicella e mpr" (ie, vaccine: varicella and mpr), and after the preprocessing, the bigram "vaccin varicell" was removed because the TFiDF weight was low. Thus the relationship between varicella and vaccine was lost and remained only the token "varicell".

The other two records were wrongly classified as negative. For one of them, the misclassification was probably due to an issue in the tokenization. In fact, an anomalous sequence of dashes ("-") and blanks lead to the token "- varicella", which was removed from the feature space, leaving no reference to the disease. The second negative misclassified record referred to a child who was vaccinated for measles, mumps, rubella, and varicella (quadrivalent vaccine). The pediatrician wrote "vaccinazione morbillo parotite rosolia varicella" (ie, vaccination, measles, mumps, rubella, varicella). The bigram "rosol varicell" (ie, "rubell varicell") was weighted 0.361 and, hence, was retained in the feature space, and was considered by all the MLTs a pattern of noninfection.

The strength of tree-based models such as LogitBoost also lies in their high scalability. In fact, their computational complexity (ie, the asymptotical time needed for a complete run) grows linearly with the sample size and quadratically with the number of features used (ie, the number of tokens considered) [37]. Assuming that the richness of the pediatric EHRs' vocabulary is limited (ie, the number of tokens reaches a plateau as data accumulates over time) an increase in computational time will only depend linearly on the number of patients.

Any attempt to use EHRs to identify patients with a specific disease would depend on the algorithm, the database, the language, and the true prevalence of the disease. As to the generalization of these models to other contexts, we hypothesize that they could also be successfully applied in public health systems with EHR charting in other languages [49].

We acknowledge that one metric (ie, sensitivity, specificity, PPV, or NPV) may be more important than another, depending on the intended use of the classification algorithm. Thus, the LogitBoost model is adequate for ascertaining varicella cases, with a preference for case identification with good sensitivity and excellent specificity.

If the aim of using MLTs is to help create a gold standard for databases, the limited agreement between the MLTs reported in Table 5 suggests that these classification algorithms are not reliable as a set of annotators.

Limitations

Some limitations must be acknowledged. First, it is acknowledged that text preprocessing is a crucial step. The way to convert free text into numbers and numbers into features is an essential step of the process and has one of the biggest impacts on the results [24]. For the same reason as before, we decided to follow a standard preprocessing procedure without searching for the best one to obtain results that are, at most, independent of human tuning.

Second, we set the number of boosting iterations as the same number of features considered. This is suboptimal in computational time because the same performance can be reached with fewer iterations [37]. Nevertheless, we aimed to reach an upper-bound value for the performance estimated in an optimal situation.

Third, the large difference in disease prevalence between the training and the validation data set should be noted. The boosting approach seems to deal with this issue in a satisfactory way, but a potential impact on model prediction could not be excluded.

Conclusions

Given their promising performance in identifying varicella cases, LogitBoost, and MLTs in general, could be effectively used for large-scale surveillance, minimizing time and cost in a scalable and reproducible manner.

Acknowledgments

The data that support the findings of this study are available from Pedianet, but restrictions apply to the availability of these data, which were used under license for this study and are not publicly available. Data are, however, available from the authors upon reasonable request and with the permission of Pedianet.

Authors' Contributions

CL, CG, and DG designed the study. CL and PB performed the analysis. CL, PB, IB, and GL wrote the manuscript. IB and DG interpreted the statistical results. GL and CG interpreted the clinical results. LT, AS, and LG handled data management.

Conflicts of Interest

None declared.

References

1. Magill SS, Dumyati G, Ray SM, Fridkin SK. Evaluating epidemiology and improving surveillance of infections associated with health care, United States. Emerg Infect Dis 2015 Sep;21(9):1537-1542 [FREE Full text] [doi: 10.3201/eid2109.150508] [Medline: 26291035]

2. Lloyd-Smith JO, Funk S, McLean AR, Riley S, Wood JL. Nine challenges in modelling the emergence of novel pathogens. Epidemics 2015 Mar;10:35-39 [FREE Full text] [doi: 10.1016/j.epidem.2014.09.002] [Medline: 25843380]
3. Sutherland SM, Kaelber DC, Downing NL, Goel VV, Longhurst CA. Electronic health record-enabled research in children using the electronic health record for clinical discovery. Pediatr Clin North Am 2016 Apr;63(2):251-268. [doi: 10.1016/j.pcl.2015.12.002] [Medline: 27017033]
4. Baracco G, Eisert S, Saavedra S, Hirsch P, Marin M, Ortega-Sanchez I. Clinical and economic impact of various strategies for varicella immunity screening and vaccination of health care personnel. Am J Infect Control 2015 Oct 01;43(10):1053-1060. [doi: 10.1016/j.ajic.2015.05.027] [Medline: 26138999]
5. Damm O, Ultsch B, Horn J, Mikolajczyk RT, Greiner W, Wichmann O. Systematic review of models assessing the economic value of routine varicella and herpes zoster vaccination in high-income countries. BMC Public Health 2015 Jun 05;15:533 [FREE Full text] [doi: 10.1186/s12889-015-1861-8] [Medline: 26041469]
6. Kawai K, Gebremeskel BG, Acosta CJ. Systematic review of incidence and complications of herpes zoster: towards a global perspective. BMJ Open 2014 Jun 10;4(6):e004833 [FREE Full text] [doi: 10.1136/bmjopen-2014-004833] [Medline: 24916088]
7. Pierik JG, Gumbs PD, Fortanier SA, Van Steenwijk PC, Postma MJ. Epidemiological characteristics and societal burden of varicella zoster virus in the Netherlands. BMC Infect Dis 2012 May 10;12:110 [FREE Full text] [doi: 10.1186/1471-2334-12-110] [Medline: 22574722]
8. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405. [doi: 10.1038/nrg3208] [Medline: 22549152]
9. Afzal Z, Schuemie MJ, van Blijderveen JC, Sen EF, Sturkenboom MC, Kors JA. Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. BMC Med Inform Decis Mak 2013 Mar 02;13:30 [FREE Full text] [doi: 10.1186/1472-6947-13-30] [Medline: 23452306]
10. Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, Hemingway H. Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning. PLoS One 2012;7(1):e30412 [FREE Full text] [doi: 10.1371/journal.pone.0030412] [Medline: 22276193]
11. Kavuluru R, Rios A, Lu Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artif Intell Med 2015 Oct;65(2):155-166 [FREE Full text] [doi: 10.1016/j.artmed.2015.04.007] [Medline: 26054428]
12. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc 2016 Sep;23(5):1007-1015 [FREE Full text] [doi: 10.1093/jamia/ocv180] [Medline: 26911811]
13. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform 2017 Jan;97:120-127 [FREE Full text] [doi: 10.1016/j.ijmedinf.2016.09.014] [Medline: 27919371]
14. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and electronic health record big data analytics for precision medicine. IEEE Trans Biomed Eng 2017 Feb;64(2):263-273 [FREE Full text] [doi: 10.1109/TBME.2016.2573285] [Medline: 27740470]
15. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22 [FREE Full text] [Medline: 20808728]
16. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). Ann Statist 2000 Apr;28(2):337-407. [doi: 10.1214/aos/1016218223]
17. Mani S, Chen Y, Arlinghaus L, Li X, Chakravarthy A, Bhave S, et al. Early prediction of the response of breast tumors to neoadjuvant chemotherapy using quantitative MRI and machine learning. AMIA Annu Symp Proc 2011;2011:868-877 [FREE Full text] [Medline: 22195145]
18. Pedianet. URL: http://www.pedianet.it/en [accessed 2019-04-09]
19. Nicolosi A, Sturkenboom M, Mannino S, Arpinelli F, Cantarutti L, Giaquinto C. The incidence of varicella: correction of a common error. Epidemiology 2003 Jan;14(1):99-102. [doi: 10.1097/00001648-200301000-00024] [Medline: 12500056]
20. Nicolosi A, Sturkenboom M, Mannino S, Arpinelli F, Cantarutti L, Giaquinto C. The incidence of varicella: correction of a common error. Epidemiology 2003 Jan;14(1):99-102. [doi: 10.1097/00001648-200301000-00024] [Medline: 12500056]
21. Cantarutti A, Donà D, Visentin F, Borgia E, Scamarcia A, Cantarutti L, Pedianet. Epidemiology of frequently occurring skin diseases in Italian children from 2006 to 2012: a retrospective, population-based study. Pediatr Dermatol 2015;32(5):668-678. [doi: 10.1111/pde.12568] [Medline: 25879514]
22. Donà D, Mozzo E, Scamarcia A, Picelli G, Villa M, Cantarutti L, et al. Community-acquired rotavirus gastroenteritis compared with adenovirus and norovirus gastroenteritis in Italian children: a Pedianet study. Int J Pediatr 2016;2016:5236243 [FREE Full text] [doi: 10.1155/2016/5236243] [Medline: 26884770]
23. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002 Mar;34(1):1-47. [doi: 10.1145/505282.505283]
24. Denny MJ, Spirling A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. Polit Anal 2018 Mar 19;26(2):168-189. [doi: 10.1017/pan.2017.44]

25. Liu M, Hu Y, Tang B. Role of text mining in early identification of potential drug safety issues. Methods Mol Biol 2014;1159:227-251. [doi: 10.1007/978-1-4939-0709-0_13] [Medline: 24788270]
26. Marafino B, Davies J, Bardach N, Dean M, Dudley RA. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. J Am Med Inform Assoc 2014;21(5):871-875 [FREE Full text] [doi: 10.1136/amiajnl-2014-002694] [Medline: 24786209]
27. Gregori D, Paola B, Soriani N, Baldi I, Lanera C. Maximizing text mining performance: the impact of pre-processing. In: JSM Proceedings, Section on Statistical Learning and Data Science. 2016 Presented at: ASA Joint Statistical Meeting; 2016; Chicago, IL p. 3265-3270.
28. Wu HC, Luk RWP, Wong KF, Kwok KL. Interpreting TF-IDF term weights as making relevance decisions. ACM Trans Inf Syst 2008 Jun 01;26(3):1-37. [doi: 10.1145/1361684.1361686]
29. Goodall CR. Data mining of massive datasets in healthcare. Journal of Computational and Graphical Statistics 1999 Sep;8(3):620-634. [doi: 10.1080/10618600.1999.10474837]
30. Jurka T. maxent: an R package for low-memory multinomial logistic regression with support for semi-automated text classification. The R Journal 2012;4(1):56. [doi: 10.32614/rj-2012-007]
31. Renner IW, Warton DI. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 2013 Mar;69(1):274-281. [doi: 10.1111/j.1541-0420.2012.01824.x] [Medline: 23379623]
32. Tuszynski J. R-project. 2019. caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc URL: https://cran.r-project.org/package=caTools
33. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics 2003 Jun 12;19(9):1061-1069. [doi: 10.1093/bioinformatics/btf867] [Medline: 12801866]
34. Freund Y, Schapire RE. Experiments with a new boosting algorithm. 340 Pine Street, Sixth Floor, San Francisco, CA: Morgan Kaufmann Publishers Inc; 1996 Jul Presented at: Thirteenth International Conference on International Conference on Machine Learning; 1996; Bari, Italy p. E URL: https://cseweb.ucsd.edu/~yfreund/papers/boostingexperiments.pdf
35. Boughorbel S, Al-Ali R, Elkum N. Model comparison for breast cancer prognosis based on clinical data. PLoS One 2016;11(1):e0146413 [FREE Full text] [doi: 10.1371/journal.pone.0146413] [Medline: 26771838]
36. Andrews P, Sleeman D, Statham P, McQuatt A, Corruble V, Jones P, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. J Neurosurg 2002 Aug;97(2):326-336. [doi: 10.3171/jns.2002.97.2.0326] [Medline: 12186460]
37. Landwehr N, Hall M, Frank E. Logistic model trees. Mach Learn 2005 May;59(1-2):161-205. [doi: 10.1007/s10994-005-0466-3]
38. Abeare S. LSU Master's Theses. 2009. Comparisons of boosted regression tree, GLM and GAM performance in the standardization of yellowfin tuna catch-rate data from the Gulf of Mexico lonline [sic] fishery URL: https://digitalcommons.lsu.edu/gradschool_theses/2880/ [accessed 2020-04-01]
39. Hastie T, Tibshirani R, Friedman J. The Elements Of Statistical Learning. Berlin, Germany: Springer Science & Business Media; 2009.
40. Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics & Data Analysis 2010 Dec;54(12):2976-2989. [doi: 10.1016/j.csda.2010.03.004]
41. Gwet K. Handbook Of Inter-rater Reliability: The Definitive Guide To Measuring The Extent Of Agreement Among Raters. Piedmont, Ca: Advanced Analytics, Llc; 2014.
42. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol 2013 Apr 29;13:61 [FREE Full text] [doi: 10.1186/1471-2288-13-61] [Medline: 23627889]
43. Zec S, Soriani N, Comoretto R, Baldi I. High agreement and high prevalence: the paradox of Cohen's kappa. Open Nurs J 2017;11:211-218 [FREE Full text] [doi: 10.2174/1874434601711010211] [Medline: 29238424]
44. R Foundation for Statistical Computing. 2016. R: A Language Environment for Statistical Computing URL: https://www.r-project.org/ [accessed 2020-04-01]
45. GitHub. mltzostercode URL: https://github.com/UBESP-DCTV/mltzostercode
46. Ross M, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform 2014 Aug 15;9:97-104 [FREE Full text] [doi: 10.15265/IY-2014-0003] [Medline: 25123728]
47. 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]
48. Xing C, Geng X, Xue H. Logistic boosting regression for label distribution learning. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV p. 4489-4497 URL: http://openaccess.thecvf.com/content_cvpr_2016/papers/Xing_Logistic_Boosting_Regression_CVPR_2016_paper.pdf [doi: 10.1109/CVPR.2016.486]
49. Lorenzoni G, Bressan S, Lanera C, Azzolina D, Da Dalt L, Gregori D. Analysis of unstructured text-based data using machine learning techniques: the case of pediatric emergency department records in Nicaragua. In: Med Care Res Rev. 2019 Apr 29 Presented at: APHA 2017 Annual Meeting & Expo; November 4-8; Atlanta, GA p. 1077558719844123. [doi: 10.1177/1077558719844123]

Abbreviations

AC1: agreement coefficient 1
DTM: document-term matrix
EHR: electronic health report
GLM: generalized linear model
GLMNet: elastic-net regularized generalized linear model
MAXENT: maximum entropy
MLT: machine learning technique
NPV: negative predicative value
PPV: positive predicative value
TF-iDF: term frequencies–inverse document frequencies

Edited by N Bruining; submitted 10.04.19; peer-reviewed by R Bajpai, M Torii, B Polepalli Ramesh; comments to author 20.06.19; revised version received 28.08.19; accepted 16.12.19; published 05.05.20.

Please cite as:
Lanera C, Berchialla P, Baldi I, Lorenzoni G, Tramontan L, Scamarcia A, Cantarutti L, Giaquinto C, Gregori D
Use of Machine Learning Techniques for Case-Detection of Varicella Zoster Using Routinely Collected Textual Ambulatory Records: Pilot Observational Study
JMIR Med Inform 2020;8(5):e14330
URL: https://medinform.jmir.org/2020/5/e14330
doi:10.2196/14330
PMID:32369038

Original Paper

Deep Learning–Based Prediction of Refractive Error Using Photorefraction Images Captured by a Smartphone: Model Development and Validation Study

Jaehyeong Chun[1*], BSc, MSc; Youngjun Kim[2*], MD; Kyoung Yoon Shin[2], MD; Sun Hyup Han[2], MD; Sei Yeul Oh[2], MD, PhD; Tae-Young Chung[2], MD, PhD; Kyung-Ah Park[2], MD, PhD; Dong Hui Lim[2,3], MD, PhD

[1]Department of Industrial and System Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
[2]Department of Ophthalmology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea
[3]Department of Digital Health, Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University, Seoul, Republic of Korea
[*]these authors contributed equally

Corresponding Author:
Dong Hui Lim, MD, PhD
Department of Ophthalmology
Samsung Medical Center
Sungkyunkwan University School of Medicine
81, Irwon-ro
Gangnam-gu
Seoul, 06351
Republic of Korea
Phone: 82 2 3410 3548
Email: ldhlse@gmail.com

Abstract

Background: Accurately predicting refractive error in children is crucial for detecting amblyopia, which can lead to permanent visual impairment, but is potentially curable if detected early. Various tools have been adopted to more easily screen a large number of patients for amblyopia risk.

Objective: For efficient screening, easy access to screening tools and an accurate prediction algorithm are the most important factors. In this study, we developed an automated deep learning–based system to predict the range of refractive error in children (mean age 4.32 years, SD 1.87 years) using 305 eccentric photorefraction images captured with a smartphone.

Methods: Photorefraction images were divided into seven classes according to their spherical values as measured by cycloplegic refraction.

Results: The trained deep learning model had an overall accuracy of 81.6%, with the following accuracies for each refractive error class: 80.0% for ≤−5.0 diopters (D), 77.8% for >−5.0 D and ≤−3.0 D, 82.0% for >−3.0 D and ≤−0.5 D, 83.3% for >−0.5 D and <+0.5 D, 82.8% for ≥+0.5 D and <+3.0 D, 79.3% for ≥+3.0 D and <+5.0 D, and 75.0% for ≥+5.0 D. These results indicate that our deep learning–based system performed sufficiently accurately.

Conclusions: This study demonstrated the potential of precise smartphone-based prediction systems for refractive error using deep learning and further yielded a robust collection of pediatric photorefraction images.

(JMIR Med Inform 2020;8(5):e16225) doi:10.2196/16225

KEYWORDS

amblyopia; cycloplegic refraction; deep learning; deep convolutional neural network; mobile phone; photorefraction; refractive error; screening

Introduction

Amblyopia is the most common cause of permanent visual impairment in children, and its worldwide prevalence is estimated to be approximately 1.6%-5% [1,2]. Refractive error is one of the leading causes of pediatric amblyopia [3]. Early detection of refractive error in children plays an important role in visual prognosis [4,5], and therefore, early pediatric screening

is recommended by the American Academy of Pediatrics, American Academy of Pediatric Ophthalmology and Strabismus (AAPOS), and European Strabismological Association and Societies [6,7].

Cycloplegic retinoscopic refraction is the standard technique for measuring refractive error. However, this method has some limitations. It is difficult to get young children to cooperate during the procedure, and advanced clinical ophthalmologic training is required to perform the test (user dependent) [2,8].

Previously, autorefractors were developed for faster and easier refraction in children. However, autorefraction presents several difficulties, including maintaining the proper position for testing and maintaining visual fixation on the target for a sufficient duration [9,10]. Photorefraction data can confirm the presence of myopia, hyperopia, astigmatism, and anisometropia by evaluating the reflection type and the position of eccentric crescent images on the pupil after projecting a light source onto the retina [11,12]. Photorefraction is simple and fast, making it convenient for use in children with poor cooperation ability, and it is suitable for screening large populations [13,14]. Several tools have been developed to meet the growing demand to perform photorefraction in clinical settings [2,15,16]. Easy availability of these tools and accurate prediction algorithms are the most important factors for ensuring efficient screening by photorefraction. Recently, deep learning algorithms have yielded innovative results in the field of medical imaging diagnostics [17]. In particular, deep convolutional neural networks [18] have been widely applied to extract essential features directly from images without human input. In ophthalmology, deep convolutional neural networks showed remarkable performance for detecting various diseases, including diabetic retinopathy [19-21], glaucoma [22,23], and retinopathy of prematurity [24]. Deep learning can also capture biological signs that are difficult for even human experts to detect, such as retinal findings from fundus images associated with cardiovascular risk [25]. However, little research has been done on the application of deep learning to refractive error prediction among children, using photorefraction images. A previous study attempted to predict the refractive error from retinal fundus images using deep learning [26], but the application was limited because the average participant age was 55 years and a specialized device was required to obtain the fundus images.

The purpose of this study was to develop an automated deep learning–based prediction system for refractive error using eccentric photorefraction images of pediatric patients captured by a smartphone. We trained our deep convolutional neural network with photorefraction images to identify various refractive error ranges. Thereafter, we comparatively evaluated its performance on our network with conventional cycloplegic retinoscopic refraction.

Methods

Study Approval

This study was performed at a single center according to the tenets of the Declaration of Helsinki. The Institutional Review Board of Samsung Medical Center (Seoul, Republic of Korea) approved this study (SMC 2017-11-114).

Participants

Patients aged 6 months to 8 years who visited the outpatient clinic for a routine ocular examination were requested to participate in this study. Written informed consent was provided by parents prior to participation. All screening tests were conducted at Samsung Medical Center between June and September 2018. The exclusion criteria were diseases that could affect light reflection, such as congenital cataracts and corneal opacity, diseases involving visual pathways or extraocular muscles, a medical history of previous ophthalmic surgery (eg, strabismus, congenital cataract, and congenital glaucoma), limited cycloplegia, and poor cooperation during study activities.

Data Collection

A total of 305 photorefraction images (191 images from 101 girls and 114 images from 63 boys) were obtained (mean age 4.32 years, SD 1.87 years). All patients underwent a complete ophthalmologic examination, including visual acuity, motility evaluation, and anterior segment evaluation. Eccentric photorefraction images were obtained using a smartphone with a 16-megapixel camera (LGM-X800K; LG Electronics Inc, Seoul, Korea) at a 1-meter distance from the front of the patient in a dark room (<15 lux). The smartphone was placed straight forward to the face of the children without angulation. All photorefraction images were acquired in the same setting (in a dark room and before the cycloplegic procedure). The smartphone's built-in flash, present next to the camera lens, was used as the light source for eccentric photorefraction, wherein light was refracted and reached the retinal surface and was then magnified and reflected. When optimal reflection was achieved, a characteristic crescent-shaped reflection appeared in the eye. A photograph of the crescent reflection was captured through LED control [13]. After acquisition of photorefraction images, 0.5% tropicamide and 0.5% phenylephrine (Tropherine; Hanmi Pharm, Seoul, Korea) were administered three times at 5-minute intervals. Cycloplegic retinoscopy and fundus examination to obtain spherical, cylindrical, cylindrical axis, and spherical equivalent values were performed between 30 and 60 minutes following the first instillation of cycloplegics, when the pupillary light reflex was eliminated. Both photorefraction and cycloplegic refraction were performed sequentially, and the ground truth for images acquired by photorefraction was labelled according to the values of cycloplegic refraction. Consequently, the result of cycloplegic refraction was provided as the ground truth for machine learning of photorefration images.

The acquired eccentric photorefraction images were divided into the following seven classes according to the spherical values measured by cycloplegic refraction: ≤−5.0 diopter (D), >−5.0 D and ≤−3.0 D, >−3.0 D and ≤−0.5 D, >−0.5 D and <+0.5 D, ≥+0.5 D and <+3.0 D, ≥+3.0 D and <+5.0 D, and ≥+5.0 D. The cutoff values of the seven classes for refractive errors were determined clinically. Among myopic refraction (minus values), −5.0 D, −3.0 D, and −0.5 D were considered as thresholds of high, moderate, and mild myopia, respectively. In other words, refractive errors ≤−5.0 D indicated high myopia, refractive

errors >−5.0 D and ≤−3.0 D indicated moderate myopia, and refractive errors >−3.0 D and ≤−0.5 D indicated mild myopia. Similarly, +0.5 D, +3.0 D, and +5.0 D were thresholds of mild, moderate, and high hyperopia, respectively, among plus values.

Image Data Preparation for Training, Validation, and Testing

Photorefraction images were processed for training our deep convolutional neural network. Initially, the images were cropped to capture the pupil. The images were resized to 224×224 pixels, and the pixel values were scaled from 0 to 1. To overcome an overfitting issue caused by an insufficiently sized training dataset, data augmentation was performed by altering brightness, saturation, hue, and contrast; adding Gaussian noise; and blurring images using Gaussian kernels. Thereafter, the image pixel values were normalized by subtracting the mean and dividing by the SD to ensure that each image had a similar data distribution and would converge faster during the training procedure.

For training, validation, and testing, we used the five-fold cross-validation approach to build a reliable deep learning model with a limited dataset. Initially, all the data were subdivided into five equal-sized folds with the same proportion of different classes in each fold. Four of the five folds were for training and validation (3.5 folds for training and 0.5 folds for validation), and one fold was for testing. After five repetitions of this process, we were able to evaluate the performance of the entire dataset because the test folds were independent of each other, and we confirmed the stability of our model for the entire dataset using the confusion matrix.

Deep Convolutional Neural Network and Training

We used a deep convolutional neural network to classify photorefraction images into the most probable class of refractive error. Among the various types of convolutional neural networks, we developed Residual Network (ResNet-18) [27] to avoid problems that occur when deep neural network depth increases, such as vanishing or exploding gradients and accuracy degradation. Residual Network addresses these issues using identity mapping with shortcut connections. The shortcut connections allow networks to skip over layers and also enable speed training. Figure 1 illustrates the overall structure of the deep learning approach we propose in this work. The basic block consists of two 3×3 convolutional layers, and the shortcut connection enables the network to learn identity mapping (Figure 2).

Because we did not have a sufficiently large training dataset, we performed transfer learning to capture low-level features, such as edge and color, without wasting image data [28]. Accordingly, pretrained parameters of Residual Network on the ImageNet [29] datasets were reused as starting points for our model. The pretrained Residual Network was available on Pytorch [30]. We then replaced the last fully connected layer to output seven predicted probabilities for each refractive error class (≤−5.0 D, >−5.0 D and ≤−3.0 D, >−3.0 D and ≤−0.5 D, >−0.5 D and <+0.5 D, ≥+0.5 D and <+3.0 D, ≥+3.0 D and <+5.0 D, and ≥+5.0 D). During the training process, the first layer was frozen, and the learning rates for the subsequent layers were increased from 1e-10 to 1e-5 to finetune our network for preventing an overfitting issue. Furthermore, we designed the loss function as a weighted sum of cross-entropy by class, wherein the weight for each class was the reciprocal of the proportion of that class's images in the training dataset. This technique was useful to achieve balanced accuracy for all classes, despite having an imbalanced training dataset. For convergence of network training, the learning rate was decayed by a factor of 0.95 every 10 epoch, and we trained the parameters of networks using stochastic gradient descent [31] with 0.9 momentum. We set the maximum training epoch as 500 and the minibatch size of training images as 16. All codes were implemented using Pytorch 1.2.0 [30]. Details of the network structure are shown in Table 1.

Figure 1. Overview of the proposed deep convolutional neural network architecture. The photorefraction image inputs pass through 17 convolutional layers and one fully connected layer, and the outputs of the network assign the probabilities for each refractive error class given the image. We also generate the localization map highlighting the important regions from the final convolutional feature maps of the layer i (i=1, 2, 3, or 4).

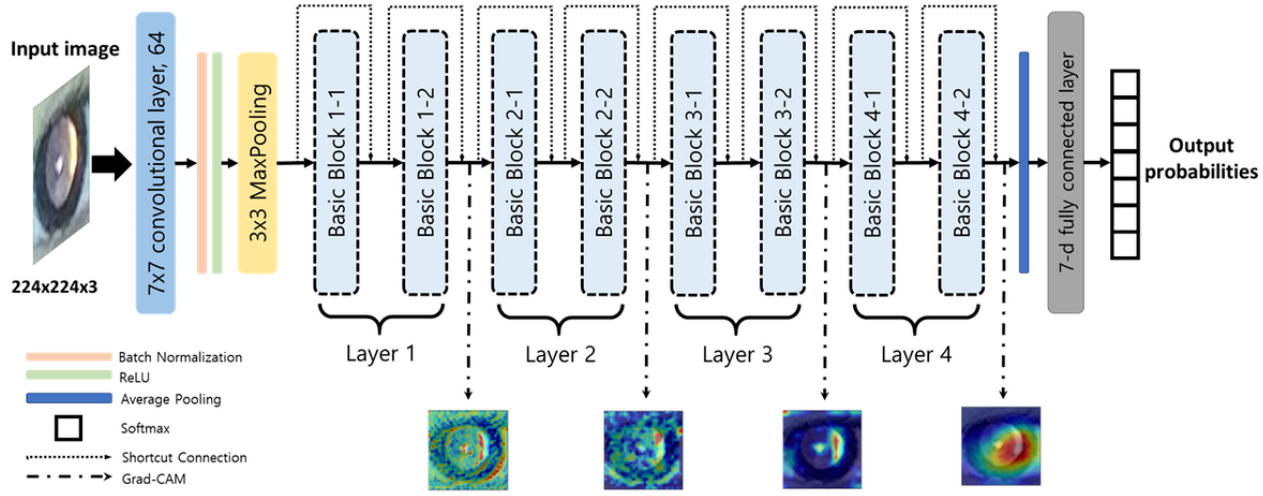


Figure 2. Structure of the basic block and the shortcut connection. The basic block consists of two 3×3 convolutional layers, two Batch Normalization layers, and a Rectified Linear Unit (ReLU) activation function. The shortcut connection adds the input vector of the basic block to the output of the basic block.

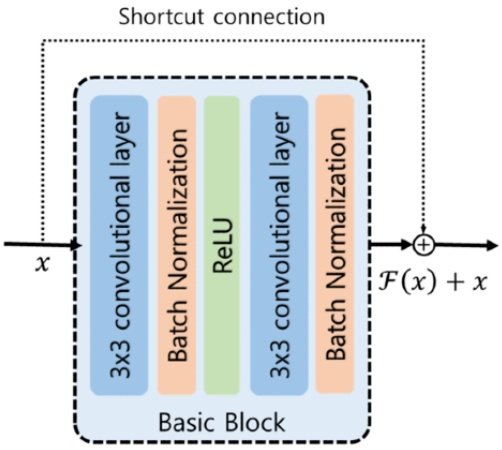


Table 1. Configuration of the deep convolutional network.

Layer type, feature map	Filters	Kernel	Stride	Padding	Learning rate
Input					
224×224×3	__[a]	—	—	—	0.0 (freeze)
Convolutional					
112×112×64	64	7×7×3	2	3	0.0 (freeze)
Batch normalization					
112×112×64	—	—	—	—	0.0 (freeze)
Max pooling					
56×56×64	1	3×3	2	1	0.0 (freeze)
Layer 1					
Basic block 1-1					
56×56×64	64	3×3×64	1	1	0.0 (freeze)
56×56×64	64	3×3×64	1	1	0.0 (freeze)
Basic block 1-2					
56×56×64	64	3×3×64	1	1	0.0 (freeze)
56×56×64	64	3×3×64	1	1	0.0 (freeze)
Layer 2					
Basic block 2-1					
28×28×128	128	3×3×64	2	1	1e-10
28×28×128	128	3×3×128	1	1	1e-10
28×28×128	128	1×1×64	2	0	1e-10
Basic block 2-2					
28×28×128	128	3×3×128	1	1	1e-10
28×28×128	128	3×3×128	1	1	1e-10
Layer 3					
Basic block 3-1					
14×14×256	256	3×3×128	2	1	1e-8
14×14×256	256	3×3×256	1	1	1e-8
14×14×256	256	1×1×128	2	0	1e-8
Basic block 3-2					
14×14×256	256	3×3×256	1	1	1e-8
14×14×256	256	3×3×256	1	1	1e-8
Layer 4					
Basic block 4-1					
7×7×512	512	3×3×256	2	1	1e-6
7×7×512	512	3×3×512	1	1	1e-6
7×7×512	512	1×1×64	2	0	1e-6
Basic block 4-2					
7×7×512	512	3×3×512	1	1	1e-6
7×7×512	512	3×3×512	1	1	1e-6
Average pooling					
1×1×512	1	7×7	7	0	—
Fully connected layer					

Layer type, feature map	Filters	Kernel	Stride	Padding	Learning rate
1×7	—	—	—	—	1e-5
Softmax					
1×7	—	—	—	—	—

[a]Not applicable.

Results

Image Dataset Demographics

A total of 305 photorefraction images from 191 girls and 114 boys were acquired. The mean age was 4.32 years (SD 1.87 years), and the median age was 4 years (range 0-8 years). The mean spherical equivalent was 0.13 D (SD 2.27 D; range −5.50 to 6.75 D), and the mean astigmatism was −1.50 D (SD 1.38 D; range −6.50 to 0 D), according to cycloplegic refraction.

According to cycloplegic refraction results, 25 photorefraction images had a refractive error ≤−5.0 D, 18 had an error >−5.0 D and ≤−3.0 D, 50 had an error >−3.0 D and ≤−0.5 D, 84 had an error >−0.5 D and <+0.5 D, 87 had an error ≥+0.5 D and <+3.0 D, 29 had an error ≥+3.0 D and <+5.0 D, and 12 had an error ≥+5.0 D. Table 2 summarizes patient demographics in detail, and examples of photorefraction images according to the refractive error class are shown in Figure 3.

Figure 3. Examples of photorefraction images from the seven different refractor error classes. A bright crescent appears in the pupillary reflex, and its size and shape indicate the diopter (D) value.

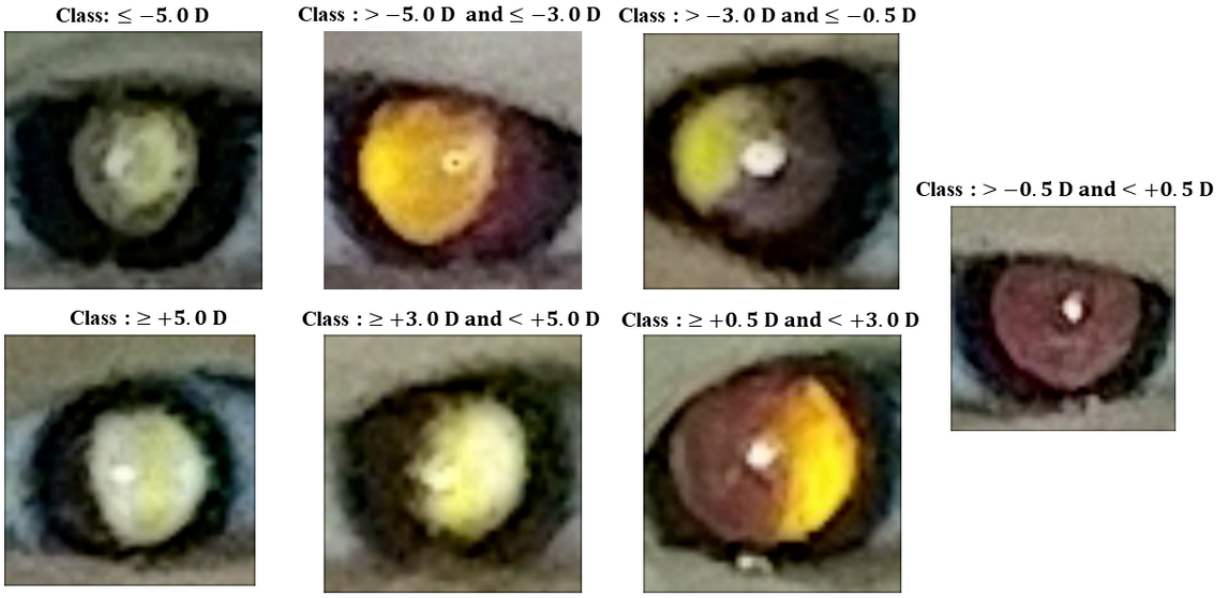


Table 2. Dataset participant demographics.

Characteristic	Value
Total images, n	305
Refractive error, n	
≤−5.0 D[a]	25
>−5.0 D and ≤−3.0 D	18
>−3.0 D and ≤−0.5 D	50
>−0.5 D and <+0.5 D	84
≥+0.5 D and <+3.0 D	87
≥+3.0 D and <+5.0 D	29
≥+5.0 D	12
Girls, n (%)	191 (62.6)
Age, mean (SD)	4.32 (1.87)

[a]D: diopters.

Performance of the Proposed Deep Convolutional Neural Network

We used five-fold cross-validation to evaluate our network's performance. Training, validation, and testing were independently iterated five times. In each iteration, there were 213 training images, 31 validation images, and 61 testing images. We chose the network with the highest validation accuracy when loss of training was saturated. Thereafter, we measured the classification accuracy of the network in the test fold. All five networks, which were established in the training phase, had an accuracy of more than 80% for each validation set. Similarly, the performances of the five testing folds were 83.6%, 80.3%, 82.0%, 78.7%, and 83.6% (Table 3).

Table 3. Results for five-fold cross-validation.

Iteration[a]	Validation accuracy (%) (N=31)	Test accuracy (%) (N=61)
First iteration	87.1	83.6
Second iteration	80.6	80.3
Third iteration	80.6	82.0
Fourth iteration	83.9	78.7
Fifth iteration	83.9	83.6
Average	83.2	81.6

[a]In each iteration, our network was trained using the rest of the validation and test dataset (213 training images).

In the five-fold test, our network had the following accuracies: 80.0% for class ≤−5.0 D, 77.8% for class >−5.0 D and ≤−3.0 D, 82.0% for class >−3.0 D and ≤−0.5 D, 83.3% for class >−0.5 D and <+0.5 D, 82.8% for class ≥+0.5 D and <+3.0 D, 79.3% for class ≥+3.0 D and <+5.0 D, and 75% for class ≥+5.0 D (Table 4). Despite the imbalanced dataset, our model achieved consistent performance for all classes.

In addition, our network maintained the stability of prediction for refractive error, as shown in the confusion matrix (Table 5). Overall, 85.7% (48/56) of total misclassifications were within one class difference and 98.2% (55/56) of total misclassifications were within two class differences.

Table 4. Performance of our deep convolutional neural network with the overall test dataset.

Class	Number	Accuracy (%)
≤−5.0 D[a]	25	80.0
>−5.0 D and ≤−3.0 D	18	77.8
>−3.0 D and ≤−0.5 D	50	82.0
>−0.5 D and <+0.5 D	84	83.3
≥+0.5 D and <+3.0 D	87	82.8
≥+3.0 D and <+5.0 D	29	79.3
≥+5.0 D	12	75.0
Total	305	81.6

[a]D: diopter.

For performance comparison, we developed the following five baseline models and calculated the performances: (1) pretrained VGG-11 [32]; (2) pretrained squeezeNet [33]; (3) Support Vector Machine (SVM) [34]; (4) Random Forest [35]; and (5) simple convolutional neural network. VGG-11 and squeezeNet were pretrained on the ImageNet [29] datasets, and their parameters were frozen, except the last four convolutional layers during training. Moreover, we designed the following two traditional machine learning approaches: SVM and Random Forest. SVM has a radial basis function kernel, 1.0 regularization parameter, and three degrees of the kernel function. Random Forests has 500 trees, the Gini index criterion, and two samples required to split an internal node. Lastly, the simple convolutional neural network has three convolutional layers with six kernels (8×8size, two strides), 16 kernels (5×5size, two strides), and 24 kernels (3×3 size, one stride), respectively; a max-pooling layer (2×2 size and two strides) after each convolutional layer; and three fully connected layers with 120, 84, and 7 hidden units, respectively, in a row at the end of the network. We evaluated the performances of the five baseline models using five-fold cross-validation, and the results of performance comparison are shown in Table 6. We confirmed that the proposed deep convolutional neural network outperformed all baseline models.

Table 5. Confusion matrix for refractive error classification of our deep convolutional neural network.

Ground truth	Predictive value							Accuracy (%)
	≤−5.0 D[a]	>−5.0 D and ≤−3.0 D	>−3.0 D and ≤−0.5 D	>−0.5 D and <+0.5 D	≥+0.5 D and <+3.0 D	≥+3.0 D and <+5.0 D	≥+5.0 D	
≤−5.0 D	20[b]	3	2	0	0	0	0	80.0
>−5.0 D and ≤−3.0 D	1	14[b]	2	0	1	0	0	77.8
>−3.0 D and ≤−0.5 D	1	4	41[b]	4	0	0	0	82.0
>−0.5 D and <+0.5 D	0	0	5	70[b]	8	1	0	83.3
≥+0.5 D and <+3.0 D	0	0	1	10	72[b]	4	0	82.8
≥+3.0 D and <+5.0 D	0	0	0	1	4	23[b]	1	79.3
≥+5.0 D	0	0	0	0	1	2	9[b]	75.0
Overall accuracy (%)	—[c]	—	—	—	—	—	—	81.6

[a]D: diopter.

[b]Number of correct predictions of our deep convolutional neural network.

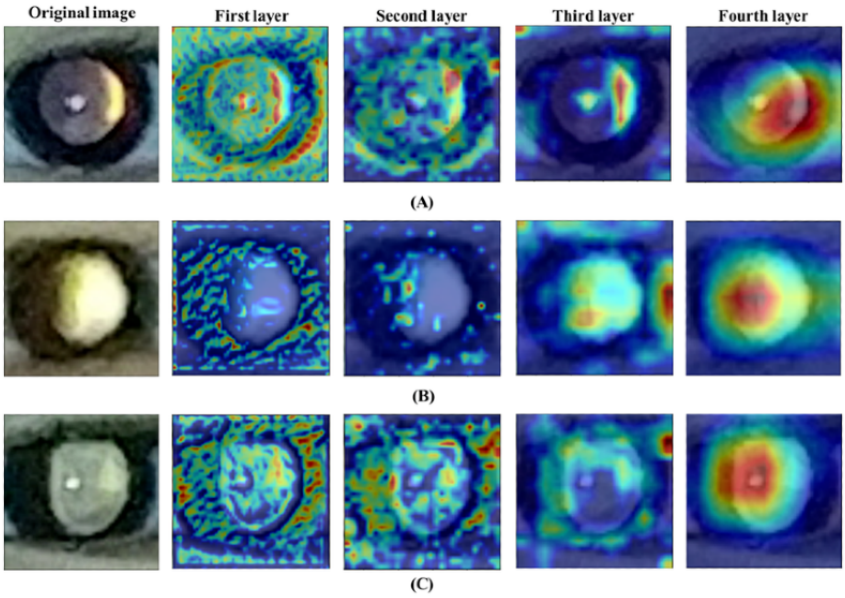
[c]Not applicable.

Table 6. Performance comparison of the proposed model and baseline models.

Model	Accuracy (%)
The proposed deep convolutional neural network	81.6
Pretrained VGG-11	70.8
Pretrained SqueezeNet	77.4
Support Vector Machine	65.2
Random Forest	62.9
Simple convolutional neural network	70.8

Additionally, we produced heatmaps using gradient-weighted class activation mapping (Grad-CAM) [36] to provide visual explanations for each screening decision. This technique is crucial for interpreting network output and validating whether the network learned meaningful features. The activation map visualizes where the network considered the critical locations to be within photorefraction images for detecting refractive error. Figure 4 shows the activated regions from four layers in the photorefraction images. Notably, we observed the heatmap from the fourth layer, which captured important features for classifying refractive error, particularly the region of the crescent in the pupil.

Figure 4. Examples of photorefraction images correctly classified by deep neural networks. (A), (B), (C) were identified as ≥+0.5 D and <+3.0 D, ≥+3.0 D and <+5.0 D, and ≥+5.0 D, respectively. The first layers captured low-level features, such as edge and color. With deeper layers, the network focused on high-level features that were regarded as important aspects for classification.



Discussion

The primary purpose of refractive error screening is the early detection of a refractive error to allow interventions that can reduce the risk of amblyopia. Early detection and treatment of refractive error can lead to better visual outcomes and reduce the prevalence and severity of amblyopia in children [4,37]. The cycloplegic refraction test has been an essential tool to accurately measure refractive error, because pediatric patients are more accommodating than adults [38]. However, young children tend not to cooperate well during the refraction test, and the test requires a skilled ophthalmic practitioner [2,8]. Additionally, the eye drops used during cycloplegia can cause side effects, such as flushing, fever, drowsiness, and red eye [39]. For these reasons, cycloplegic refraction is not suitable for large screening of refractive error and amblyopia [12]. Currently, smartphones are ubiquitous devices that allow physicians and other related medical professionals to overcome common diagnostic barriers in many clinical settings [40]. A photorefraction screening test using a smartphone is an easy and effective way to screen most young children. The photorefractive method is simple and takes no longer than a second to test both eyes simultaneously. The test requires minimal space (just a meter of distance between the subject and the testing device) and removes the need for cycloplegia, thereby greatly reducing side effects and testing time. Moreover, it does not require expert knowledge or experience to perform [6]. These advantages make the photorefractive method ideal for measuring refractive error, especially for poorly cooperative young children.

Several studies have compared the accuracy of photoscreeners for detecting various amblyopia risk factors [40-42]. One study evaluated a smartphone photoscreening application (GoCheckKids) and reported 76% sensitivity and 67.2% specificity [15] for detecting amblyopia risk factors using the 2013 AAPOS guidelines. Because we evaluated the accuracy of predicting refractive errors and not amblyopia risk factors, we were limited in our ability to directly compare the performance of our method against that of GoCheckKids. Instead, our deep convolutional neural network achieved satisfactory accuracy for predicting categories of refractive error using only a small image dataset. The results showed the potential for developing precise smartphone-based prediction systems for refractive error using deep learning. With further collection of pediatric photorefraction image data, more precise prediction of refractive error and effective detection of amblyopia would be possible.

This study compared refractive error estimation with precycloplegic photorefraction images and cycloplegic refraction. The results showed consistent measurements between the two methods. Dubious results regarding estimation of refractive error using photorefractors have been uncovered by previous studies [12,14,42]. Erdurmus et al reported that noncycloplegic photorefraction (Plusoptix CR03; PlusoptiX GMBH, Nurnberg, Germany) tended to overestimate negative refraction in children, resulting in overdiagnosis of myopia (−0.70 D) [12]. Lim et al reported similar results and showed that refractive error measured by a photorefractor without cycloplegia (PlusoptiX S09; PlusoptiX GmbH) tended to be more myopic compared with cycloplegic refractive error [42]. On the other hand, Schimizek et al claimed that noncycloplegic refraction using a photorefractometer (Power Refractor; PlusoptiX GmbH) resulted in underestimation of spherical equivalents owing to uncontrolled accommodation [14]. Another study showed that cycloplegic refraction results and photorefractor Plusoptix S08 (Plusoptix GmbH, Nurnberg, Germany) results were similar [2]. In this study, photorefraction results without cycloplegia showed reasonable agreement with cycloplegic refraction, suggesting that our deep learning–based system achieved considerably accurate performance under noncycloplegic conditions.

This study has several limitations. First, manifest refraction was not performed in all subjects. Since photorefractive refraction tests were performed without the use of a cycloplegic agent, useful information might have been obtained if the number of manifest refraction results without cycloplegia were enough to compare with photorefraction data in the same patient. Second, the number of photorefraction images was relatively small and the model could only predict a range of refractive errors (not a specific value). Third, all children involved in the study were Korean. Thus, a trained model using the eyes of Korean children may not be applicable to the eyes of pediatric patients having different ethnicities [43,44]. Future studies with more patients of multiple ethnicities and a greater range of refractive errors would be beneficial for providing a more precise clinical perspective.

In conclusion, this study showed that our deep learning–based system successfully yielded accurate and precise refractive measurements. This further demonstrates the potential for developing simplified smartphone-based prediction systems for refractive error using deep learning with large-scale collection of pediatric photorefraction images from patients with various ages and refractive errors.

Acknowledgments

This research was supported by a National Research Foundation of Korea grant funded by the Government of Korea's Ministry of Education (NRF-2018R1D1A1A02045884; Seoul, Korea), which was received by Dong Hui Lim, and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI19C0577), which was received by Dong Hui Lim.

Authors' Contributions

DHL designed the study. JC, YK, KYS, DHL, and K-AP analyzed and interpreted the clinical data. JC and YK wrote the submitted manuscript draft. TYC, SYO, SHH, DHL, and KAP reviewed the design, the results, and the submitted draft. JC and YK contributed equally to the work as cofirst authors. DHL and KAP are the corresponding authors for this study. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. Simons K. Amblyopia characterization, treatment, and prophylaxis. Surv Ophthalmol 2005 Mar;50(2):123-166. [doi: 10.1016/j.survophthal.2004.12.005] [Medline: 15749306]
2. Demirci G, Arslan B, Özsütçü M, Eliaçık M, Gulkilik G. Comparison of photorefraction, autorefractometry and retinoscopy in children. Int Ophthalmol 2014 Aug 10;34(4):739-746. [doi: 10.1007/s10792-013-9864-x] [Medline: 24114503]
3. Miller JM, Dobson V, Harvey EM, Sherrill DL. Cost-efficient vision screening for astigmatism in native american preschool children. Invest Ophthalmol Vis Sci 2003 Sep 01;44(9):3756-3763. [doi: 10.1167/iovs.02-0970] [Medline: 12939288]
4. Cotter SA, Pediatric Eye Disease Investigator Group, Edwards AR, Wallace DK, Beck RW, Arnold RW, et al. Treatment of anisometropic amblyopia in children with refractive correction. Ophthalmology 2006 Jun;113(6):895-903 [FREE Full text] [doi: 10.1016/j.ophtha.2006.01.068] [Medline: 16751032]
5. U.S. Preventive Services Task Force. Screening for visual impairment in children younger than age 5 years: recommendation statement. Ann Fam Med 2004 May 01;2(3):263-266 [FREE Full text] [doi: 10.1370/afm.193] [Medline: 15209205]
6. Schimitzek T, Haase W. Efficiency of a video-autorefractometer used as a screening device for amblyogenic factors. Graefes Arch Clin Exp Ophthalmol 2002 Sep 27;240(9):710-716. [doi: 10.1007/s00417-002-0524-5] [Medline: 12271366]
7. American Academy of Ophthalmology Pediatric Ophthalmology/Strabismus Panel. Preferred Practice Pattern Guidelines. Pediatric Eye Evaluations. American Academy of Ophthalmology 2007.
8. Safir A. Retinoscopy. Int Ophthalmol Clin 1971;11(1):115-129. [doi: 10.1097/00004397-197101110-00008] [Medline: 5129703]
9. Prabakaran S, Dirani M, Chia A, Gazzard G, Fan Q, Leo SW, et al. Cycloplegic refraction in preschool children: comparisons between the hand-held autorefractor, table-mounted autorefractor and retinoscopy. Ophthalmic Physiol Opt 2009 Jul;29(4):422-426. [doi: 10.1111/j.1475-1313.2008.00616.x] [Medline: 19523087]
10. La TY, Oh JR. Reliability of Refractive Measurement by Hand-held Autorefractor. J Korean Ophthalmol Soc 2002:2241-2245.
11. Donahue SP, Arnold RW, Ruben JB. Preschool vision screening: what should we be detecting and how should we report it? Uniform guidelines for reporting results of preschool vision screening studies. Journal of American Association for Pediatric Ophthalmology and Strabismus 2003 Oct;7(5):314-316. [doi: 10.1016/s1091-8531(03)00182-4]
12. Erdurmus M, Yagci R, Karadag R, Durmus M. A comparison of photorefraction and retinoscopy in children. J AAPOS 2007 Dec;11(6):606-611. [doi: 10.1016/j.jaapos.2007.04.006] [Medline: 17588794]
13. Cole TD. Multimeridian Photorefraction: A technique for the detection of visual defects in infants and preverbal children. Johns Hopkins APL Technical Digest 1991:166-175.

14. Schimitzek T, Lagrèze WA. Accuracy of a new photo-refractometer in young and adult patients. Graefes Arch Clin Exp Ophthalmol 2005 Jul 14;243(7):637-645. [doi: 10.1007/s00417-004-1056-y] [Medline: 15650858]
15. Peterseim MM, Rhodes RS, Patel RN, Wilson ME, Edmondson LE, Logan SA, et al. Effectiveness of the GoCheck Kids Vision Screener in Detecting Amblyopia Risk Factors. Am J Ophthalmol 2018 Mar;187:87-91. [doi: 10.1016/j.ajo.2017.12.020] [Medline: 29305313]
16. Forcina BD, Peterseim MM, Wilson ME, Cheeseman EW, Feldman S, Marzolf AL, et al. Performance of the Spot Vision Screener in Children Younger Than 3 Years of Age. Am J Ophthalmol 2017 Jun;178:79-83 [FREE Full text] [doi: 10.1016/j.ajo.2017.03.014] [Medline: 28336401]
17. Litjens G, Kooi T, Bejnordi BE, Setio AA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017 Dec;42:60-88. [doi: 10.1016/j.media.2017.07.005] [Medline: 28778026]
18. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Commun ACM. 2017 May 24 Presented at: Proceedings of the 25th International Conference on Neural Information Processing Systems; 2012; Lake Tahoe, NV, USA p. 84-90. [doi: 10.1145/3065386]
19. Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. Invest Ophthalmol Vis Sci 2016 Oct 01;57(13):5200-5206. [doi: 10.1167/iovs.16-19964] [Medline: 27701631]
20. 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]
21. Gargeya R, Leng T. Automated Identification of Diabetic Retinopathy Using Deep Learning. Ophthalmology 2017 Jul;124(7):962-969. [doi: 10.1016/j.ophtha.2017.02.008] [Medline: 28359545]
22. Shibata N, Tanito M, Mitsuhashi K, Fujino Y, Matsuura M, Murata H, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. Sci Rep 2018 Oct 02;8(1):14665 [FREE Full text] [doi: 10.1038/s41598-018-33013-w] [Medline: 30279554]
23. Li Z, He Y, Keel S, Meng W, Chang RT, He M. Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. Ophthalmology 2018 Aug;125(8):1199-1206. [doi: 10.1016/j.ophtha.2018.01.023] [Medline: 29506863]
24. Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RV, ImagingInformatics in Retinopathy of Prematurity (i-ROP) Research Consortium. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA Ophthalmol 2018 Jul 01;136(7):803-810 [FREE Full text] [doi: 10.1001/jamaophthalmol.2018.1934] [Medline: 29801159]
25. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar 19;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
26. Varadarajan AV, Poplin R, Blumer K, Angermueller C, Ledsam J, Chopra R, et al. Deep Learning for Predicting Refractive Error From Retinal Fundus Images. Invest Ophthalmol Vis Sci 2018 Jun 01;59(7):2861-2868. [doi: 10.1167/iovs.18-23887] [Medline: 30025129]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016 Presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 27-30, 2016; Las Vegas, NV, USA URL: https://www.cvfoundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf [doi: 10.1109/cvpr.2016.90]
28. Yosinski J, Clune J, Bengio Y, Lipson H. Advances in neural information processing systems. 2014. How transferable are features in deep neural networks? URL: https://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf [accessed 2019-07-01]
29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 2015 Apr 11;115(3):211-252. [doi: 10.1007/s11263-015-0816-y]
30. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press; 2019:8024-8035.
31. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958 Nov;65(6):386-408. [doi: 10.1037/h0042519] [Medline: 13602029]
32. Simonyan K, Zisserman A. arXiv. 2014. Very deep convolutional networks for large-scale image recognition URL: https://arxiv.org/pdf/1409.1556.pdf [accessed 2020-04-09]
33. Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. arXiv. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size URL: https://arxiv.org/pdf/1602.07360.pdf [accessed 2020-04-09]
34. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge, United Kingdom: Cambridge University Press; 2000.
35. Breiman L. Random Forests. Machine Learning 2001;45:5-32. [doi: 10.1023/A:1010933404324]
36. Selvaraju RR, 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. 2017 Presented

at: 2017 IEEE International Conference on Computer Vision (ICCV); October 22-29, 2017; Venice, Italy URL: http://openaccess.thecvf.com/content_ICCV_2017/papers/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.pdf [doi: 10.1109/iccv.2017.74]
37. Scheiman MM, Hertle RW, Beck RW, Edwards AR, Birch E, Cotter SA, Pediatric Eye Disease Investigator Group. Randomized trial of treatment of amblyopia in children aged 7 to 17 years. Arch Ophthalmol 2005 Apr 01;123(4):437-447. [doi: 10.1001/archopht.123.4.437] [Medline: 15824215]
38. Morgan IG, Iribarren R, Fotouhi A, Grzybowski A. Cycloplegic refraction is the gold standard for epidemiological studies. Acta Ophthalmol 2015 Sep 18;93(6):581-585 [FREE Full text] [doi: 10.1111/aos.12642] [Medline: 25597549]
39. Wakayama A, Nishina S, Miki A, Utsumi T, Sugasawa J, Hayashi T, et al. Incidence of side effects of topical atropine sulfate and cyclopentolate hydrochloride for cycloplegia in Japanese children: a multicenter study. Jpn J Ophthalmol 2018 Sep 25;62(5):531-536. [doi: 10.1007/s10384-018-0612-7] [Medline: 30046935]
40. Arnold RW, O'Neil JW, Cooper KL, Silbert DI, Donahue SP. Evaluation of a smartphone photoscreening app to detect refractive amblyopia risk factors in children aged 1–6 years. OPTH 2018 Aug;12:1533-1537. [doi: 10.2147/opth.s171935]
41. Arnold RW, Armitage MD. Performance of four new photoscreeners on pediatric patients with high risk amblyopia. J Pediatr Ophthalmol Strabismus 2014 Jan 03;51(1):46-52. [doi: 10.3928/01913913-20131223-02] [Medline: 24369683]
42. Lim JH, Bae GH, Shin SJ. Reliability and Usefulness of Refractive Measurements by PlusoptiX S09 in Children. J Korean Ophthalmol Soc 2014;55(7):1071. [doi: 10.3341/jkos.2014.55.7.1071]
43. Sravani NG, Nilagiri VK, Bharadwaj SR. Photorefraction estimates of refractive power varies with the ethnic origin of human eyes. Sci Rep 2015 Jan 23;5(1):7976 [FREE Full text] [doi: 10.1038/srep07976] [Medline: 25613165]
44. Bharadwaj SR, Sravani NG, Little J, Narasaiah A, Wong V, Woodburn R, et al. Empirical variability in the calibration of slope-based eccentric photorefraction. J Opt Soc Am A 2013 Apr 19;30(5):923. [doi: 10.1364/josaa.30.000923]

Abbreviations

AAPOS: American Academy of Pediatric Ophthalmology and Strabismus
SVM: Support Vector Machine

Edited by G Eysenbach; submitted 12.09.19; peer-reviewed by G Lim, J Hamer, S Kim, A Davoudi, M Banf; comments to author 09.10.19; revised version received 03.03.20; accepted 20.03.20; published 05.05.20.

Please cite as:
Chun J, Kim Y, Shin KY, Han SH, Oh SY, Chung TY, Park KA, Lim DH
Deep Learning–Based Prediction of Refractive Error Using Photorefraction Images Captured by a Smartphone: Model Development and Validation Study
JMIR Med Inform 2020;8(5):e16225
URL: https://medinform.jmir.org/2020/5/e16225
doi:10.2196/16225
PMID:32369035

Original Paper

Deep Learning Neural Networks to Predict Serious Complications After Bariatric Surgery: Analysis of Scandinavian Obesity Surgery Registry Data

Yang Cao[1], PhD; Scott Montgomery[1,2,3], PhD; Johan Ottosson[4], MD, PhD; Erik Näslund[5], MD, PhD; Erik Stenberg[4], MD, PhD

[1]Clinical Epidemiology and Biostatistics, School of Medical Sciences, Örebro University, Örebro, Sweden

[2]Clinical Epidemiology Division, Department of Medicine, Karolinska Institutet, Stockholm, Sweden

[3]Department of Epidemiology and Public Health, University College London, London, United Kingdom

[4]Department of Surgery, Faculty of Medicine and Health, Örebro University, Örebro, Sweden

[5]Division of Surgery, Department of Clinical Sciences, Danderyd Hospital, Karolinska Institutet, Stockholm, Sweden

Corresponding Author:
Yang Cao, PhD
Clinical Epidemiology and Biostatistics
School of Medical Sciences
Örebro University
Campus USÖ
Örebro, 70182
Sweden
Phone: 46 196026236
Email: yang.cao@oru.se

Abstract

Background: Obesity is one of today's most visible public health problems worldwide. Although modern bariatric surgery is ostensibly considered safe, serious complications and mortality still occur in some patients.

Objective: This study aimed to explore whether serious postoperative complications of bariatric surgery recorded in a national quality registry can be predicted preoperatively using deep learning methods.

Methods: Patients who were registered in the Scandinavian Obesity Surgery Registry (SOReg) between 2010 and 2015 were included in this study. The patients who underwent a bariatric procedure between 2010 and 2014 were used as training data, and those who underwent a bariatric procedure in 2015 were used as test data. Postoperative complications were graded according to the Clavien-Dindo classification, and complications requiring intervention under general anesthesia or resulting in organ failure or death were considered serious. Three supervised deep learning neural networks were applied and compared in our study: multilayer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN). The synthetic minority oversampling technique (SMOTE) was used to artificially augment the patients with serious complications. The performances of the neural networks were evaluated using accuracy, sensitivity, specificity, Matthews correlation coefficient, and area under the receiver operating characteristic curve.

Results: In total, 37,811 and 6250 patients were used as the training data and test data, with incidence rates of serious complication of 3.2% (1220/37,811) and 3.0% (188/6250), respectively. When trained using the SMOTE data, the MLP appeared to have a desirable performance, with an area under curve (AUC) of 0.84 (95% CI 0.83-0.85). However, its performance was low for the test data, with an AUC of 0.54 (95% CI 0.53-0.55). The performance of CNN was similar to that of MLP. It generated AUCs of 0.79 (95% CI 0.78-0.80) and 0.57 (95% CI 0.59-0.61) for the SMOTE data and test data, respectively. Compared with the MLP and CNN, the RNN showed worse performance, with AUCs of 0.65 (95% CI 0.64-0.66) and 0.55 (95% CI 0.53-0.57) for the SMOTE data and test data, respectively.

Conclusions: MLP and CNN showed improved, but limited, ability for predicting the postoperative serious complications after bariatric surgery in the Scandinavian Obesity Surgery Registry data. However, the overfitting issue is still apparent and needs to be overcome by incorporating intra- and perioperative information.

(JMIR Med Inform 2020;8(5):e15992) doi:10.2196/15992

KEYWORDS

projections and predictions; deep learning; computational neural networks; bariatric surgery; postoperative complications

Introduction

Background

Obesity is one of today's most important public health problems worldwide. With no changes in the current trends, the estimated prevalence of severe obesity (BMI greater than 35 kg/m^2) will reach 9% for women and 6% for men within a few years [1]. Obesity is associated with an increased risk of several conditions and diseases, such as type 2 diabetes, heart disease, and many more, and imposes a major growing threat for global public health [2]. It is a serious chronic condition that should be prevented and treated as early as possible [3]. Although medical weight management and pharmacotherapy are effective options, modern bariatric surgery offers one of the best chances for long-term weight loss and the resolution of comorbidity risk [4].

Although modern bariatric surgery is considered to be ostensibly safe, serious complications and mortality still occur in some patients [5-7]. Thus, preoperative risk assessment is one of the most important components of surgical decision making. Numerous studies have attempted to predict the risk for complications after bariatric surgery. Some studies developed new models based on national databases [5-9], and other studies applied the obesity surgery mortality risk score, although its accuracy for prediction is still unclear [7,10-14]. In recent years, the potential of addressing public health challenges and advancing medical research through the increasing amount of information regarding symptoms, diseases, and treatments, in parallel with the challenges inherent in working with such sources, are being recognized [15]. A variety of machine learning (ML) methods, including artificial neural networks [16], decision trees [17], Bayesian networks [18], and support vector machines [19], have been widely applied with the aim of detecting key features of the patient conditions and modeling the disease progression after treatment from complex health information and medical datasets. The application of different ML methods in feature selection and classification in multidimensional heterogeneous data can provide promising tools for inference in medical practices [20,21]. These highly nonlinear approaches have been utilized in medical research for the development of predictive models, resulting in effective and accurate decision making [22-24].

In our previous studies, conventional statistical models [8] and ML methods [9] were used to predict the likelihood of serious complication after bariatric surgery. Although some potential risk factors, such as revision surgery, age, lower BMI, larger waist circumference (WC), and dyspepsia, were associated with a higher risk for serious postoperative complications by the multivariate logistic regression model, the sensitivity of the model for prediction was quite low (<0.01) [8]. When comparing 29 ML algorithms, we found that overfitting was still the overwhelming problem even though some algorithms showed both high accuracy >0.95 and an acceptable area under curve (AUC) >0.90 for the training data [9]. Despite these unfavorable aspects, our study suggests that deep learning neural networks (DLNNs) have the potential to improve the predictive capability and deserve further investigation.

Although there is increasing evidence that the use of ML methods can improve our understanding of postoperative progression of bariatric surgery [25-30], few studies have used DLNNs to predict the prognosis after bariatric surgery, and validation is needed to select a proper method in clinical practice.

Objectives

The aim of this study was to examine whether serious postoperative complications of bariatric surgery can be predicted preoperatively using DLNNs based on the information available from a national quality registry. We used the data from the Scandinavian Obesity Surgery Registry (SOReg) to examine the performance of 3 widely used DLNNs.

Methods

Patients and Features

The SOReg covers virtually all bariatric surgical procedures performed in Sweden since 2010 [31]. Patients who were registered in the SOReg between 2010 and 2015 were included in this study. Information for the patients who underwent a bariatric procedure between 2010 and 2014 was used as training data, and information from those in 2015 was used as test data. Postoperative complications were graded according to the Clavien-Dindo classification, and complications requiring intervention under general anesthesia or resulting in organ failure or death were considered serious (ie, grade 3b or higher) [32]. The primary outcome was serious complications occurring within the first 30 days after bariatric surgery. Details of the data have been described elsewhere [8,9]. Briefly, 37,811 and 6250 patients were used as the training data and test data, with incidence rates of serious complication of 3.2% (1220/37,811) and 3.0% (188/6250), respectively. In general, the patients with and without serious complication were balanced in baseline demographic characteristics and comorbidity in the 2 datasets, except that the patients with serious complications were a little older (mean 42.9 vs 41.2 years; P<.001) and had greater WCs (mean 126.2 vs 123.2 cm; P=.009) compared with those without serious complications in the test dataset [9]. Except for the outcome variable, 16 features of the patients were used for ML, including 5 continuous features (age, hemoglobin A_{1c} [HbA_{1c}], BMI, WC, and operation year) and 11 dichotomous features (sex; sleep apnea; hypertension; diabetes; dyslipidemia; dyspepsia; depression; musculoskeletal pain; previous venous thromboembolism; revisional surgery; and the outcome, serious postoperative complications).

The Regional Ethics Committee in Stockholm approved the study (approval number: 2013/535-31/5).

Deep Learning Neural Networks

Three supervised DLNNs were applied and compared in our study, comprising multilayer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN) models. For the MLP model, we used 4 dense layers and 2 dropout layers. The initial computation units for the dense layers were set to 15, 64, 64, and 128, and dropout rate was set to 0.5 for the 2 dropout layers (Multimedia Appendix 1). The rectified linear unit (relu) activation function was used for the 3 dense layers, and the sigmoid activation function was used for the last dense layer. The binary cross-entropy loss function and the root mean square propagation optimizer were used when compiling the model [33].

In the initial CNN, we used a 7-layer model with 2 one-dimensional (1D) convolution layers (with 10 filters for each), 2 1D max pooling layers, 1 flatten layer, and 2 dense layers (with 1000 computation units). The relu activation function was used for the 2 1D convolution layers and the first dense layers, and the sigmoid activation function was used for the last dense layer. The binary cross-entropy loss function and the adaptive moment estimation (Adam) optimizer were used when compiling the model (Multimedia Appendix 2) [34].

In view of the temporal feature of the data, we also used the RNN for prediction. To minimize computation time, the initial model only included 1 long short-term memory (LSTM) layer and 1 dense layer. The initial dimensionality of the LSTM layer was set to 32. To tackle overfitting, we randomly dropped out inputs and recurrent connections in the LSTM layer to break happenstance correlations in the training data that the layer was exposed to. The dropout rates for inputs and recurrent connections were set to 0.2. The activation functions for input connection and recurrent connection were hyperbolic tangent and hard sigmoid, respectively. The activation function for the dense layer was sigmoid. The binary cross-entropy loss function and the Adam optimizer were used when compiling the model.

Feature Scaling

For the training data, the binary features were converted into dummy variables, and the continuous features were standardized to have mean 0 and SD 1 before they enter the model. For the test data, the continuous features were standardized using the corresponding means and standardizations from the training data. HbA_{1c} was log transformed before standardization because of its asymmetrical distribution. In sensitivity analysis, the normalizer and min-max scaler were also used to evaluate the influence of scalers on the models' performance.

Data Augmentation

As the incidence rate of serious complications is very low (only 3.2%), the extreme imbalance would result in serious bias in the performance metrics [35]. Therefore, we used the synthetic minority oversampling technique (SMOTE) to artificially augment the proportion of patients with serious complications. SMOTE generates a synthetic instance by interpolating the *m* instances (for a given integer value *m*) of the minority class that lies close enough to each other to achieve the desired ratio between the majority and minority classes [36]. In our study, a SMOTE dataset with a 1:1 ratio between the patients with and without serious complications was generated and used for training.

Performance Metrics

The performances of the three neural networks were evaluated using accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC) [37], and area under the receiver operating characteristic (ROC) curve. Terminology and derivations of the metrics are given in detail elsewhere [9]. A successful prediction model was defined as with an AUC greater than 0.7 [38,39].

Validation During Model Training

To find optimal high-level parameters (such as the number, size, and type of layers in the networks) and lower-level parameters (such as the number of epochs, choice of loss function and activation function, and optimization procedure) in the DLNN models, the K-fold cross-validation method was used during the training phase. K-fold cross-validation is currently considered as a minimum requirement to handle the problems such as overfitting when applying only 1 single dataset in ML [40]. In this study, we split the training data into 5 partitions, instantiated 5 identical models, and trained each one on 4 partitions while evaluating the remaining partition. We then computed the average performance metrics over the 5 folds. In the end, the choice of the parameters was a compromise between the neural network's performance and computation time: the model with a larger ratio of AUC to logarithmic computation time or no significant difference ($\Delta AUC \leq 0.01$) found between the models' performance. An example of parameters selection by grid searching for MLP model is given in Multimedia Appendix 3.

Software and Hardware

The descriptive and inferential statistical analyses were performed using Stata 15.1 (StataCorp LLC, College Station). The DLNN models were achieved using packages scikit-learn 0.19.1 and Keras 2.1.6 in Python 3.6 (Python Software Foundation). The 95% CI of AUC was calculated using the package pROC in R 3.61 (R Foundation for Statistical Computing).

All the computation was conducted using a computer with the 64-bit Windows 7 Enterprise operating system (Service Pack 1), Intel Core TM i5-4210U CPU of 2.40 GHz, and 16.0 GB installed random access memory.

Results

Overview of the Performance of the 3 Deep Learning Neutral Networks

The incidence of serious complications after bariatric surgery in our study was 3.2%, which is similar to other studies [12,41]. The 3 DLNNs showed quite similar performance for our original training data, with specificity=1.00, sensitivity=0, and AUC≤0.6 (Table 1). Although the models' specificity dropped when trained using SMOTE data, the sensitivity increased significantly from 0 to 0.97 in the MLP model and 0.70 in the CNN model (Table 1), and AUC also achieved an acceptable level (>0.7). The finding confirms our previous assumption that DLNNs trained by SMOTE data might have better performance in

predicting serious complications after bariatric surgery [9]. However, the performance of the 3 DLNNs in the test data was still low; the highest AUC was only 0.23 for the MLP trained by the SMOTE data (Table 1). MCC measures indicate that the MLP trained by the SMOTE data showed promising prediction (MCC=0.44) for the training data; however, the performance of the 3 DLNNs was only slightly better than random prediction (MCC=0.02, 0.03, and 0.05 for MLP, CNN, and RNN, respectively) for the test data (Table 1).

Table 1. Performance metrics of the models.

Model	Training data					Test data				
	Accuracy	Specificity	Sensitivity	MCC[a]	AUC[b] (95% CI)	Accuracy	Specificity	Sensitivity	MCC	AUC (95% CI)
MLP[c]	0.97	1.00	0.00	0.00	0.60 (0.59-0.61)	0.97	1.00	0.00	0.00	0.57 (0.55-0.59)
MLP[d]	0.68	0.39	0.97	0.44	0.84 (0.83-0.85)	0.84	0.82	0.23	0.02	0.54 (0.53-0.55)
CNN[e]	0.97	1.00	0.00	0.00	0.58 (0.56-0.60)	0.97	1.00	0.00	0.00	0.55 (0.54-0.56)
CNN[d]	0.63	0.56	0.70	0.26	0.79 (0.78-0.80)	0.95	0.97	0.06	0.03	0.57 (0.59-0.61)
RNN[f]	0.97	1.00	0.00	0.00	0.58 (0.57-0.59)	0.97	1.00	0.00	0.00	0.56 (0.55-0.57)
RNN[d]	0.58	0.66	0.49	0.15	0.65 (0.64-0.66)	0.91	0.93	0.14	0.05	0.55 (0.53-0.57)

[a]MCC: Matthews correlation coefficient.

[b]AUC: area under curve.

[c]MLP: multilayer perceptron.

[d]Trained using synthetic minority oversampling technique data.

[e]CNN: convolutional neural network.

[f]RNN: recurrent neural network.

Performance of Multilayer Perception

There were myriad combinations of high- and low-level parameters used during model training, and most of them resulted in constant performance after given values. Therefore, we only show the trend of the MLP model's accuracy with number of epochs for model training while keeping other parameters unchanged in Figure 1. When learning from the original data, the accuracy almost did not change along with the number of epochs, which was a constant value 0.968 (Figure 1, left panel). The reason is that the incidence rate of serious complications was only 3.2%; therefore, although the model always predicted a patient as having a serious complication, it achieved high accuracy (>0.96), whereas in the SMOTE data where the numbers of patients with and without serious complications are equal, the choice of number of epochs shows a significant influence on accuracy. When the epochs are less than 20, the accuracy is smaller than 0.8, and it approximates to 0.85 when epochs are greater than 80 and remains almost constant afterward (Figure 1, right panel). As the computing time is proportional to the number of epochs, we selected epochs 80 for model training.

The performance of the MLP was not optimal for the original training data and test data. The AUCs were barely higher than a random guess, that is, 0.5, which were 0.60 (95% CI 0.59-0.61) and 0.57 (95% CI 0.55-0.59) for the training data and test data, respectively (Figure 2, left panel). When trained using the SMOTE data, the performance of the MLP improved notably, with an AUC of 0.84 (95% CI 0.83-0.85). However, its performance was still low for the test data, with an AUC of 0.54 (95% CI 0.53-0.55; Figure 2, left panel).

The performance of MLP was significantly influenced by the number of computation units in the SMOTE data but not in the test data. For example, when the computation units of the first layer ranged from 4 to 500, the AUC increased rapidly from 0.55 to 0.80. Within the range from 500 to 1000, the AUC increased slowly from 0.80 to 0.85 and kept fluctuating around 0.85 afterward (Figure 3). However, the AUC kept fluctuating around 0.55 in the test data no matter how many units were used (Figure 3).

Figure 1. Change of accuracy with the number of epochs in multilayer perceptron. MLP: multilayer perceptron; SMOTE: synthetic minority oversampling technique.

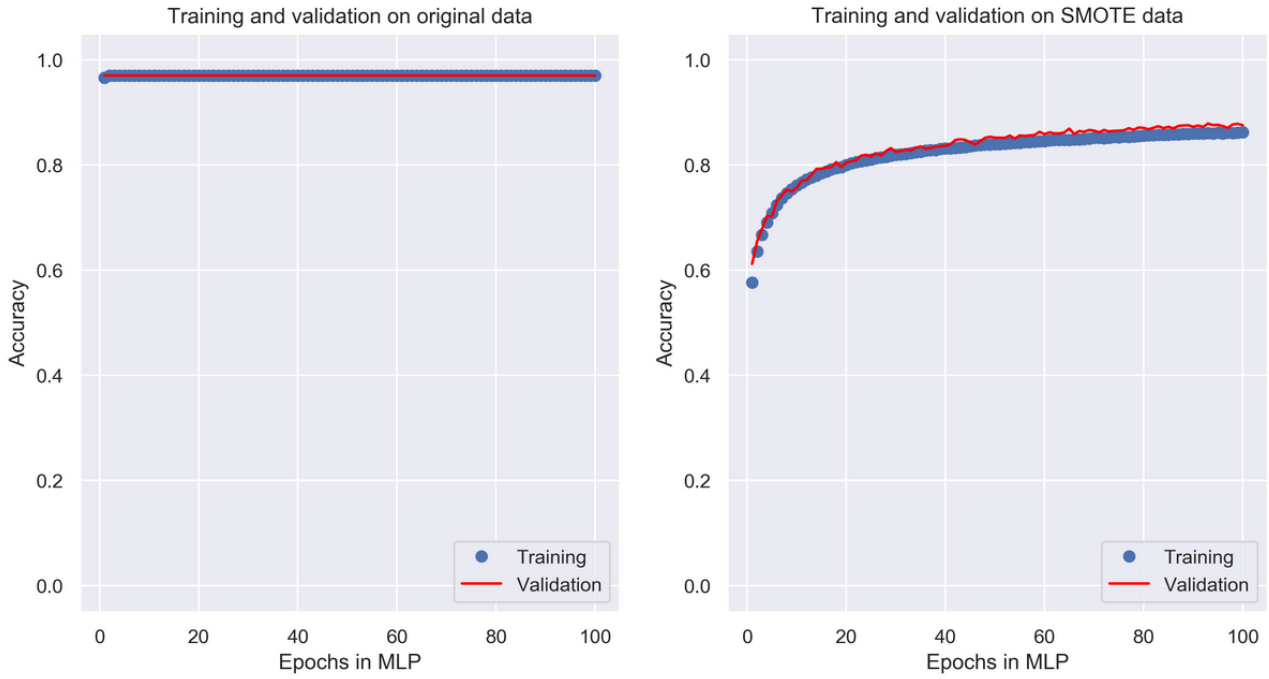


Figure 2. Area under curve of multilayer perceptron with initial setting. MLP: multilayer perceptron.

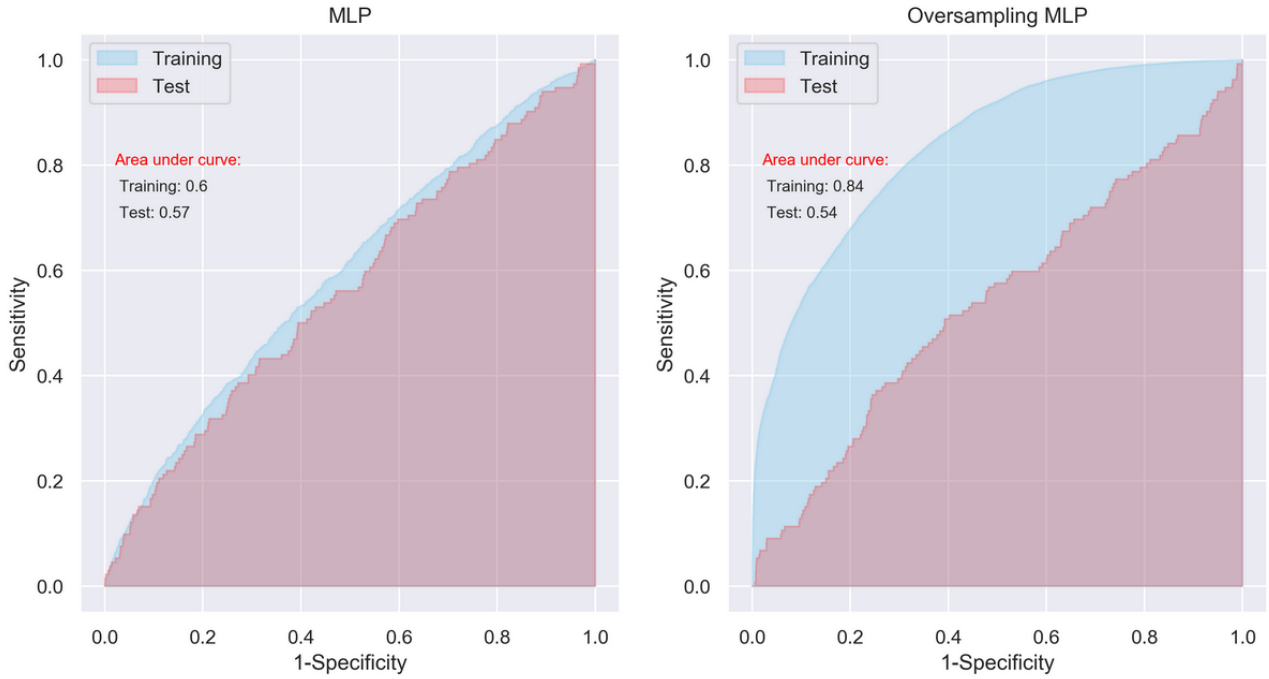
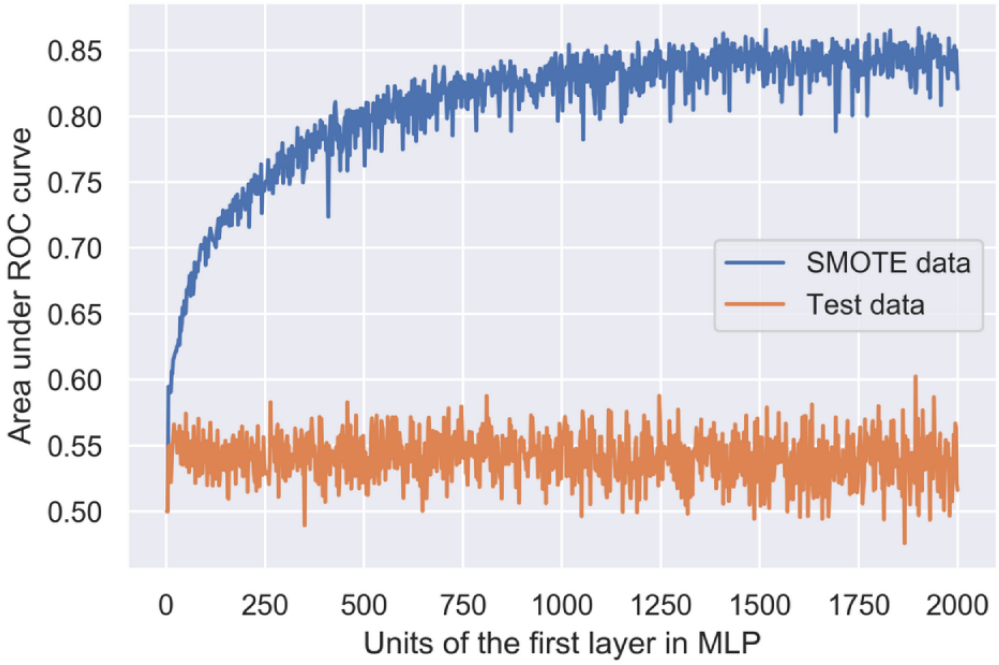


Figure 3. Performance of multilayer perceptron using the synthetic minority oversampling technique and test data with different numbers of computation units in the first hidden layer. MLP: multilayer perceptron; ROC: receiver operating characteristic; SMOTE: synthetic minority oversampling technique.



Performance of Convolutional Neutral Network

The performance of CNN appeared to be similar to that of MLP. The AUCs were 0.58 (95% CI 0.56-0.60) and 0.55 (95% CI 0.54-0.56) for the training data and test data, respectively (Figure 4, left panel). When trained using the SMOTE data, the AUCs were 0.79 (95% CI 0.78-0.80) and 0.57 (95% CI 0.59-0.61), respectively (Figure 4, right panel). Again, although the model's performance seems to be improved significantly after training by the artificially balanced SMOTE data, its performance on the unseen test data still appears low.

The number of output filters in the convolution (or the dimensionality of the output space) has a significant influence on the CNN model's performance in the SMOTE data but not in the training data and test data. The AUC of CNN increased rapidly from 0.63 to 0.80 when we set the number of filters from 5 to 50. However, the larger number of filters contributes no further improvement (Figure 5). The CNN model trained by the SMOTE data always gave an AUC around 0.52 in the test data (Figure 5).

Figure 4. Area under curve of convolutional neural network with initial setting. CNN: convolutional neural network.

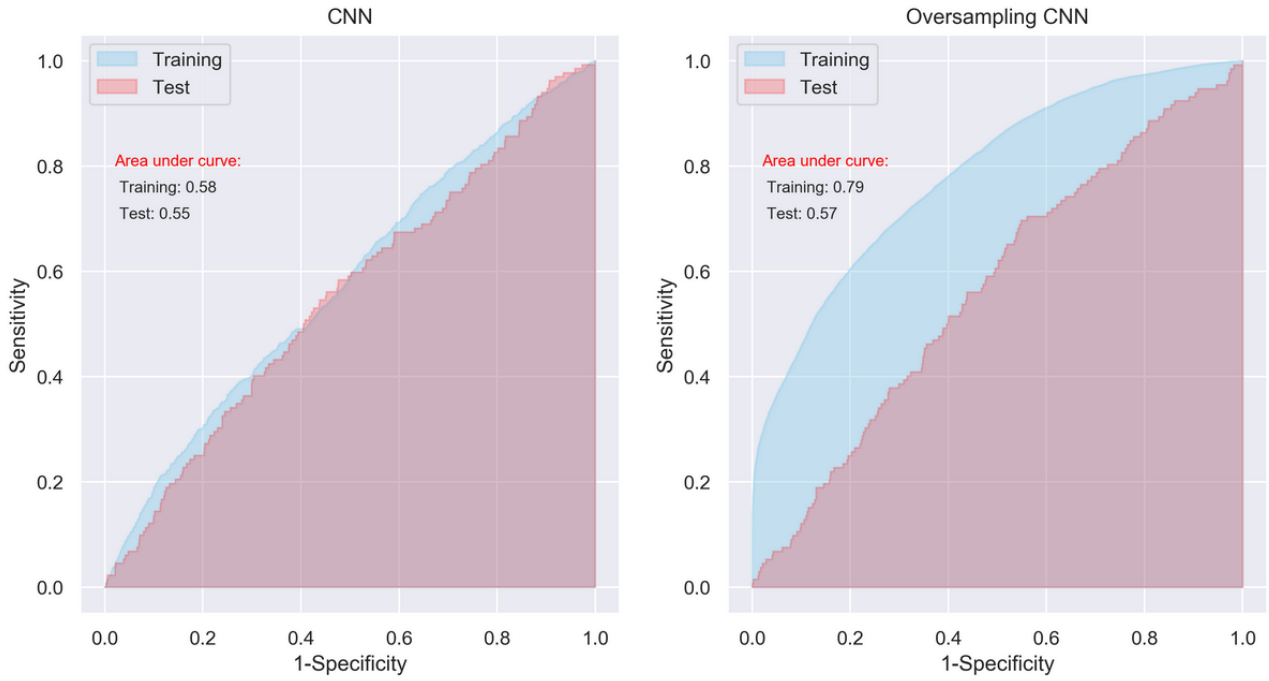
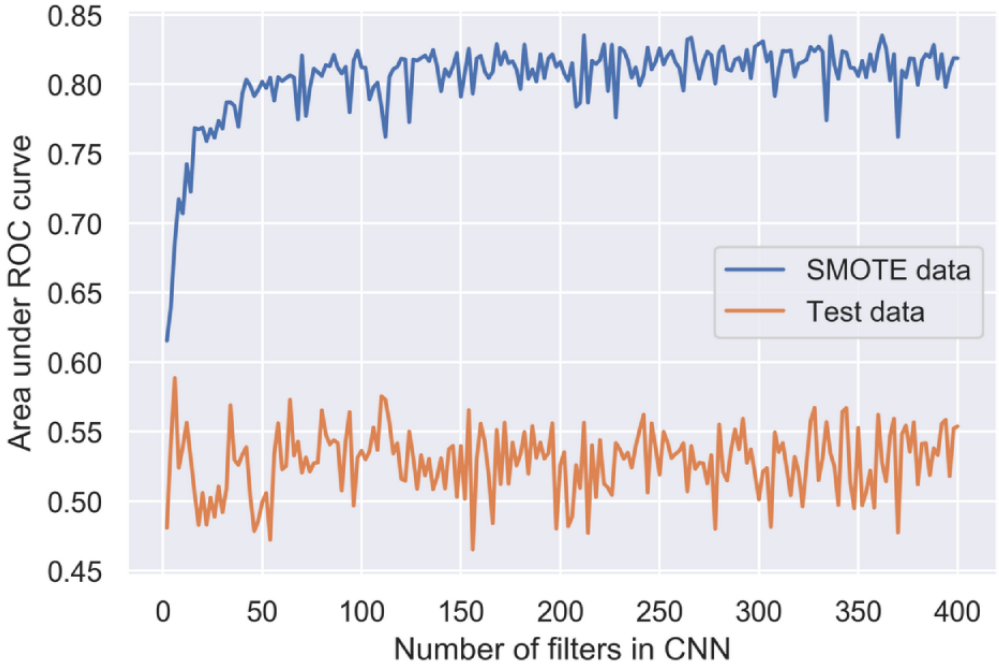


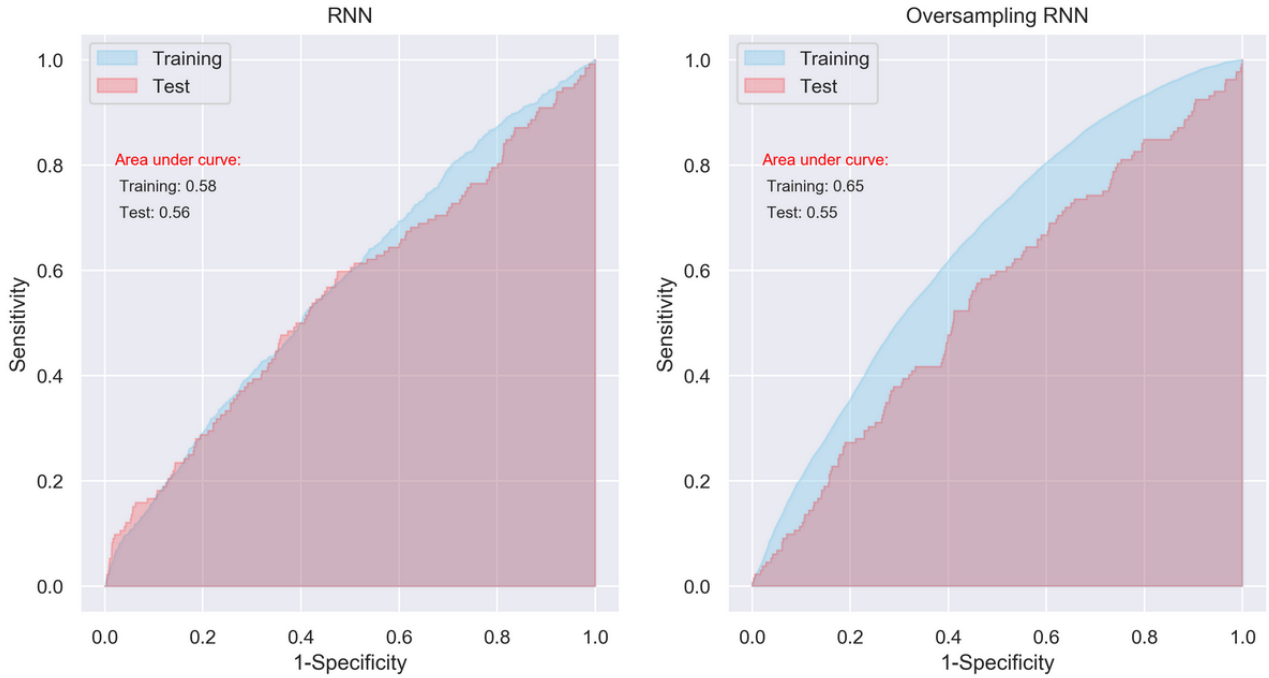
Figure 5. Performance of convolutional neural network using the synthetic minority oversampling technique and test data with different numbers of filters. CNN: convolutional neural network; ROC: receiver operating characteristic; SMOTE: synthetic minority oversampling technique.



Performance of Recurrent Neutral Network

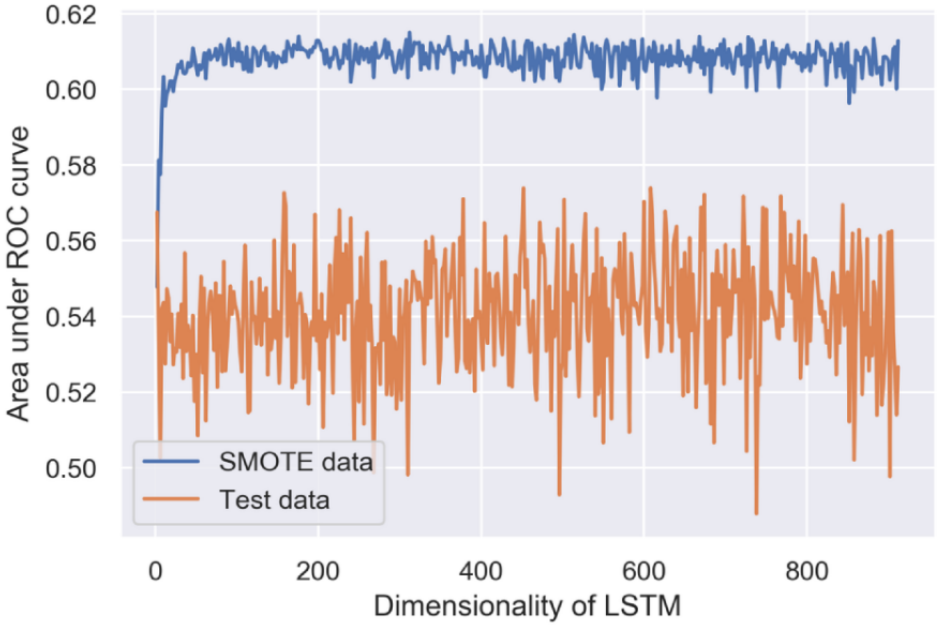
Compared with the MLP and CNN, the RNN showed even worse performance. AUCs of RNN for the original training data and test data were 0.58 (95% CI 0.57-0.59) and 0.56 (95% CI 0.55-0.57), respectively (Figure 6, left panel). For the SMOTE data, the AUC was only 0.65 (95% CI 0.64-0.66; Figure 6, right panel), which was significantly lower than those derived from MLP (AUC=0.83) and CNN (AUC=0.81). The AUC of RNN trained by the SMOTE data was only 0.55 (95% CI 0.53-0.57) for the test data.

Figure 6. Area under curve of recurrent neural network with initial setting. RNN: recurrent neural network.



The performance of the RNN model was influenced by the dimensionality of the LSTM layer. The AUC changed from 0.50 to 0.60 rapidly when the dimensionality grew from 2 to 20 and kept fluctuating around 0.61 afterward (Figure 7). Although other hyperparameters, such as kernel initializer and regularizer, also had an influence on the RNN's performance, their impacts were not as notable as the dimensionality of layer.

Figure 7. Performance of recurrent neural network using the synthetic minority oversampling technique and test data with different dimensionalities of long short-term memory. LSTM: long short-term memory; ROC: receiver operating characteristic; SMOTE: synthetic minority oversampling technique.



Sensitivity Analysis and Computing Time

In the sensitivity analysis, we tried different scalers and optimizers in data preparation and model compiling, and we tried thousands of combinations of hyperparameters for each model using the exhaustive grid search method [42]. Although they showed more or less influence on the models' performance, the influence was negligible compared with the exponentially increased computing time. Therefore, we only show the results of the model with the optimal hyperparameters in the figures above.

The computing time for the models was largely dependent on the number of DLNN layers and hyperparameter settings of the layers, number of epochs and batch size for training, and obviously software and hardware used. In our study, with the model structures and hyperparameters described above, the running time ranged from 82 seconds for the MLP model (computational units=64, epochs=80, batch size=128, and trained by original data) to more than 10 hours for the CNN model (filters=400, epochs=100, batch size=128, and trained by SMOTE data with cross-validation and grid search) on our computer.

Discussion

Principal Findings

Several studies have explored using ML methods to predict the risks after bariatric surgery. Razzaghi et al [27] evaluated 6 of the most popular classification methods to predict 4 common outcomes (diabetes, angina, heart failure, and stroke) using 11,636 patients from the Premier Healthcare Database of the United States. The study also applied the SMOTE technique to handle the imbalance issue in the data, and the results indicate that random forest and bagging methods outperform other methods [27]. However, the study did not test methods using outer unseen data. Therefore, the real performance of the methods is questionable. Thomas et al [28] predicted the long-term weight status after bariatric surgery in 478 patients using 8 neural networks. Their neural networks yielded an AUC of 0.77 to 0.78 in predicting weight loss success. However, the types of the neural networks used were not reported. It seems as if the authors only used 1 neural network but with different variables as input. Pedersen et al [25] used neural networks integrating clinical and genomic biomarkers for 268 patients to rank factors involved in type 2 diabetes remission after bariatric surgery, and Hayes et al [26] used the decision tree and the Naive Bayes to establish independent predictors for the resolution of type 2 diabetes in 130 patients. However, the sample sizes of both studies seem too small for nonlinear ML algorithms; therefore, models might only have a high internal validity but not external validity [43]. In our previous study, we trained and compared 29 basic ML algorithms using information from 37,811 patients to predict serious complications after bariatric surgery. Although several ensemble algorithms, such as random forest, gradient regression tree, and bagging k-nearest neighbor, showed favorable performance, the overfitting problem was apparent [9].

In this study, we applied and compared 3 DLNN models for predicting serious complications after bariatric surgery. MLP is the classical type of neural network, which consists of multiple layers of computational units. The layers are interconnected in a feedforward way, where the information moves only forward, that is, from input nodes, through hidden nodes and to output nodes, and the connections between the nodes do not form a cycle [44]. CNN is a regularized version of MLP, which was inspired by biological processes where the connectivity pattern between neurons resembles the organization of the animal's visual cortex [45]. Although not specifically developed for nonimage data, CNN may achieve state-of-the-art results for classification prediction problems using time series data or

sequence input data. The CNN input is traditionally two-dimensional but can also be changed to be 1D, allowing it to develop an internal representation of a 1D sequence. RNN is designed to work with sequence prediction problems and traditionally difficult to train, where connections between nodes form a directed graph along a temporal sequence, which allows it to exhibit temporal dynamic behavior. Unlike feedforward neural networks, RNN can use its internal state (memory) to process sequences of inputs. In effect, an RNN is a type of neural network that has an internal loop. It loops over time steps, and at each time step, it considers its current state at *t* and input at *t* and combines them to obtain the output at *t* [46]. RNN is traditionally difficult to train, but the LSTM network overcomes the problems of training a recurrent network and, in turn, has been perhaps the most successful and widely applied. Therefore, we adopted the LSTM network in this study. Regarding the choice of the number of layers in DLNNs, there is no universally agreed upon threshold, but most researchers in the field agree that DLNN has multiple nonlinear layers with a credit assignment path (CAP) >2, and Schmidhuber [44] considers CAP >10 to be very deep learning. To address a specific real-world predictive modeling problem, in general, we cannot analytically calculate the number of layers or the number of nodes in a DLNN and have to use systematic experimentation to discover what works best for our specific dataset.

Although the results from the MLP and CNN models seem promising in the SMOTE training data, the overfitting problem still exists, which was reflected in the poor performance of the 3 models in the test data (see Table 1 and the left panels in Figures 2, 4, and 6). It means that although we have identified potential risk factors related to serious complication after bariatric surgery at the population level [8], using current data available to predict whether an individual patient has a serious complication after bariatric surgery is still far from clinically applicable. Thus, despite using the most promising methods of ML, these results support a previous review of standard statistical methods for the prediction of complications in bariatric surgery, where models based only on factors known before surgery were insufficient to predict postoperative complications [47]. The main reason for this insufficiency is likely to be that all such methods are missing information on intraoperative adverse events, surgical experience, and perioperative optimization of patients, which are well-known important risk factors for adverse postoperative outcome [7,47-49].

We also noticed that the RNN performed worse than MLP and RNN for our data. The possible reason might be that the sequential pattern or temporal trend in our data cannot be represented by the features currently available in our data, or there is no dependency between the patients or events in the time-series. Even if the trend can be captured by the RNN, it might be weak, and the past status contributed noise rather than information to current status.

Although increasing the number of computational units in the layers or adding more layers may increase the model's capacity, the trade-off between computational expensiveness and representational power is seen everywhere in ML. Limited by the computing power, we tried to avoid complicated networks such as applying multiple RNN layers or combining CNN and RNN, but it deserves investigation in the future with data having more variables and apparent temporal trend.

Advantages and Limitations

Compared with previous studies, there are several advantages in our study. First, we used DLNNs rather than traditional ML techniques. The biggest advantage of DLNNs is that they try to learn high-level features from data in an incremental manner. They need less human domain expertise for hard-core feature extraction [50]. In contrast, in traditional ML techniques, most of the applied features have to be identified by domain experts to reduce the complexity of the data and make patterns more visible to learning algorithms to work [44]. Second, the study is based on a national quality register with extensive coverage (97%) of the target population, with a very high follow-up rate for the studied outcome. Therefore, on the one hand, the selection bias is minimized in the study, and the much larger sample size may ensure the external validity of the nonlinear ML algorithms. Third, we conducted different types of sensitivity analyses for feature scaling, hyperparameters optimization, and model compiling during data training, which ensure the efficiency and internal validity of our models. However, we also have to admit that there are still some limitations in our study. First, because of the low predictive ability of the features available in SOReg in terms of the Nagelkerke R^2 and AUC [8,9], we failed to diminish overfitting of the DLNN models. We hope to solve this problem by incorporating extra variables on perioperative care in the future. Including these factors is likely to improve the predictive ability; however, these models would not allow guidance in the preoperative setting. Second, although the DLNN models are efficient and able to formulate an adequate solution to the particular question, they are highly specialized to the specific domain, and retraining is usually necessary for the questions that do not pertain to the identical domain [51]. For example, if we want to predict a specific serious complication such as pulmonary embolism after bariatric surgery, we have to modify the layers and readjust hyperparameters in the model because the original models were not trained differentially for the different outcomes. Third, DL requires a large amount of computing power. The high-performance hardwire such as the multicore graphics processing unit is usually needed. It is time consuming and costly, and we have to give up some of the more complicated models because of extreme time inefficiency and leave them for future investigation when more efficient algorithms or more powerful hardware become available.

Conclusions

Compared with the results from our previous study using traditional ML algorithms to predict the postoperative serious complication after bariatric surgery using SOReg data, the MLP and CNN showed improved, but limited, predictive ability, which deserves further investigation. The overfitting issue is still apparent and needs to be overcome by incorporating more patient features, for example, intra- and perioperative information, from other data resources.

Acknowledgments

YC's work was supported by Örebro Region County Council (OLL-864441). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Structure of the MLP model. MLP: multilayer perceptron.
[PNG File , 121 KB - medinform_v8i5e15992_app1.png]

Multimedia Appendix 2
Structure of the CNN model. CNN: convolutional neural network.
[PNG File , 135 KB - medinform_v8i5e15992_app2.png]

Multimedia Appendix 3
Performance of the first 175 MLP models with different computation units, epochs, and batch sizes. In general, performance of the models (measured as AUC) increased with more computation units and epochs, and decreased with larger batch sizes. Although the performance increased with the model's complexity, the efficiency (measured as AUC divided by logarithmic computing time) decreased. MLP: multilayer perceptron; AUC: area under the curve.
[PNG File , 178 KB - medinform_v8i5e15992_app3.png]

References

1. NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. Lancet 2016 Apr 2;387(10026):1377-1396 [FREE Full text] [doi: 10.1016/S0140-6736(16)30054-X] [Medline: 27115820]
2. Kopelman PG, Caterson ID, Dietz WH. Clinical Obesity in Adults and Children. New York: John Wiley & Sons; 2009.
3. Adams TD, Davidson LE, Litwin SE, Kolotkin RL, LaMonte MJ, Pendleton RC, et al. Health benefits of gastric bypass surgery after 6 years. J Am Med Assoc 2012 Sep 19;308(11):1122-1131 [FREE Full text] [doi: 10.1001/2012.jama.11164] [Medline: 22990271]
4. Sjöström L, Lindroos A, Peltonen M, Torgerson J, Bouchard C, Carlsson B, Swedish Obese Subjects Study Scientific Group. Lifestyle, diabetes, and cardiovascular risk factors 10 years after bariatric surgery. N Engl J Med 2004 Dec 23;351(26):2683-2693. [doi: 10.1056/NEJMoa035622] [Medline: 15616203]
5. Finks JF, Kole KL, Yenumula PR, English WJ, Krause KR, Carlin AM, Michigan Bariatric Surgery Collaborative, from the Center for Healthcare Outcomes and Policy. Predicting risk for serious complications with bariatric surgery: results from the Michigan Bariatric Surgery Collaborative. Ann Surg 2011 Oct;254(4):633-640. [doi: 10.1097/SLA.0b013e318230058c] [Medline: 21897200]
6. Gupta PK, Franck C, Miller WJ, Gupta H, Forse RA. Development and validation of a bariatric surgery morbidity risk calculator using the prospective, multicenter NSQIP dataset. J Am Coll Surg 2011 Mar;212(3):301-309. [doi: 10.1016/j.jamcollsurg.2010.11.003] [Medline: 21247780]
7. Stenberg E, Szabo E, Agren G, Näslund E, Boman L, Bylund A, Scandinavian Obesity Surgery Registry Study Group. Early complications after laparoscopic gastric bypass surgery: results from the Scandinavian Obesity Surgery Registry. Ann Surg 2014 Dec;260(6):1040-1047. [doi: 10.1097/SLA.0000000000000431] [Medline: 24374541]
8. Stenberg E, Cao Y, Szabo E, Näslund E, Näslund I, Ottosson J. Risk prediction model for severe postoperative complication in bariatric surgery. Obes Surg 2018 Jul;28(7):1869-1875 [FREE Full text] [doi: 10.1007/s11695-017-3099-2] [Medline: 29330654]
9. Cao Y, Fang X, Ottosson J, Näslund E, Stenberg E. A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. J Clin Med 2019 May 12;8(5):pii: E668 [FREE Full text] [doi: 10.3390/jcm8050668] [Medline: 31083643]
10. DeMaria EJ, Portenier D, Wolfe L. Obesity surgery mortality risk score: proposal for a clinically useful score to predict mortality risk in patients undergoing gastric bypass. Surg Obes Relat Dis 2007;3(2):134-140. [doi: 10.1016/j.soard.2007.01.005] [Medline: 17386394]
11. Efthimiou E, Court O, Sampalis J, Christou N. Validation of Obesity Surgery Mortality Risk Score in patients undergoing gastric bypass in a Canadian center. Surg Obes Relat Dis 2009;5(6):643-647. [doi: 10.1016/j.soard.2009.08.010] [Medline: 19837010]

12. Sarela AI, Dexter SP, McMahon MJ. Use of the obesity surgery mortality risk score to predict complications of laparoscopic bariatric surgery. Obes Surg 2011 Nov;21(11):1698-1703. [doi: 10.1007/s11695-011-0379-0] [Medline: 21399971]
13. Geubbels N, de Brauw LM, Acherman YI, van de Laar AW, Wouters MW, Bruin SC. The preceding surgeon factor in bariatric surgery: a positive influence on the learning curve of subsequent surgeons. Obes Surg 2015 Aug;25(8):1417-1424. [doi: 10.1007/s11695-014-1538-x] [Medline: 25511752]
14. Lorente L, Ramón JM, Vidal P, Goday A, Parri A, Lanzarini E, et al. Obesity surgery mortality risk score for the prediction of complications after laparoscopic bariatric surgery. Cir Esp 2014 May;92(5):316-323. [doi: 10.1016/j.ciresp.2013.09.014] [Medline: 24361099]
15. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of predicting health care utilization via web search behavior: a data-driven analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
16. Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. Neural Netw 2006 May;19(4):408-415. [doi: 10.1016/j.neunet.2005.10.007] [Medline: 16483741]
17. Esteban C, Arostegui I, Moraza J, Aburto M, Quintana JM, Pérez-Izquierdo J, et al. Development of a decision tree to assess the severity and prognosis of stable COPD. Eur Respir J 2011 Dec;38(6):1294-1300 [FREE Full text] [doi: 10.1183/09031936.00189010] [Medline: 21565913]
18. Verduijn M, Peek N, Rosseel PM, de Jonge E, de Mol BA. Prognostic Bayesian networks I: rationale, learning procedure, and clinical use. J Biomed Inform 2007 Dec;40(6):609-618 [FREE Full text] [doi: 10.1016/j.jbi.2007.07.003] [Medline: 17704008]
19. Barakat NH, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed 2010 Jul;14(4):1114-1120. [doi: 10.1109/TITB.2009.2039485] [Medline: 20071261]
20. Liu H, Motoda H. Computational Methods of Feature Selection. New York: CRC Press; 2007.
21. Lee H, Yoon SB, Yang S, Kim WH, Ryu H, Jung C, et al. Prediction of acute kidney injury after liver transplantation: machine learning approaches vs logistic regression model. J Clin Med 2018 Nov 8;7(11):pii: E428 [FREE Full text] [doi: 10.3390/jcm7110428] [Medline: 30413107]
22. Ali AR. Emerj - Artificial Intelligence Research and Insight. 2017. Deep Learning in Oncology – Applications in Fighting Cancer URL: https://emerj.com/ai-sector-overviews/deep-learning-in-oncology/ [accessed 2018-06-01]
23. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol 2017 Dec;2(4):230-243 [FREE Full text] [doi: 10.1136/svn-2017-000101] [Medline: 29507784]
24. Tran BX, Vu GT, Ha GH, Vuong Q, Ho M, Vuong T, et al. Global evolution of research in artificial intelligence in health and medicine: a bibliometric study. J Clin Med 2019 Mar 14;8(3):pii: E360 [FREE Full text] [doi: 10.3390/jcm8030360] [Medline: 30875745]
25. Pedersen HK, Gudmundsdottir V, Pedersen MK, Brorsson C, Brunak S, Gupta R. Ranking factors involved in diabetes remission after bariatric surgery using machine-learning integrating clinical and genomic biomarkers. NPJ Genom Med 2016;1:16035 [FREE Full text] [doi: 10.1038/npjgenmed.2016.35] [Medline: 29263820]
26. Hayes MT, Hunt LA, Foo J, Tychinskaya Y, Stubbs RS. A model for predicting the resolution of type 2 diabetes in severely obese subjects following Roux-en Y gastric bypass surgery. Obes Surg 2011 Jul;21(7):910-916. [doi: 10.1007/s11695-011-0370-9] [Medline: 21336560]
27. Razzaghi T, Safro I, Ewing J, Sadrfaridpour E, Scott JD. Predictive models for bariatric surgery risks with imbalanced medical datasets. Ann Oper Res 2019;280(1-2):1-18. [doi: 10.1007/s10479-019-03156-8]
28. Thomas DM, Kuiper P, Zaveri H, Surve A, Cottam DR. Neural networks to predict long-term bariatric surgery outcomes. Bariatric Times 2017;14(12):14-17 [FREE Full text]
29. Piaggi P, Lippi C, Fierabracci P, Maffei M, Calderone A, Mauri M, et al. Artificial neural networks in the outcome prediction of adjustable gastric banding in obese women. PLoS One 2010 Oct 27;5(10):e13624 [FREE Full text] [doi: 10.1371/journal.pone.0013624] [Medline: 21048960]
30. Ehlers AP, Roy SB, Khor S, Mandagani P, Maria M, Alfonso-Cristancho R, et al. Improved risk prediction following surgery using machine learning algorithms. EGEMS (Wash DC) 2017 Apr 20;5(2):3 [FREE Full text] [doi: 10.13063/2327-9214.1278] [Medline: 29881747]
31. Hedenbro JL, Näslund E, Boman L, Lundegårdh G, Bylund A, Ekelund M, et al. Formation of the Scandinavian Obesity Surgery Registry, SOReg. Obes Surg 2015 Oct;25(10):1893-1900. [doi: 10.1007/s11695-015-1619-5] [Medline: 25703826]
32. Dindo D, Demartines N, Clavien P. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Ann Surg 2004 Aug;240(2):205-213. [doi: 10.1097/01.sla.0000133083.54934.ae] [Medline: 15273542]
33. Kurbiel T, Khaleghian S. arXiv preprint. 2017. Training of Deep Neural Networks based on Distance Measures using RMSProp URL: https://arxiv.org/pdf/1708.01911.pdf [accessed 2020-03-04]
34. Kingma DP, Ba JL. arXiv preprint. 2014. Adam: A Method for Stochastic Optimization URL: https://arxiv.org/pdf/1412.6980.pdf [accessed 2020-03-04]
35. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor Newsl 2004 Jun;6(1):20-29. [doi: 10.1145/1007730.1007735]

36. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321-357. [doi: 10.1613/jair.953]
37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975 Oct 20;405(2):442-451. [doi: 10.1016/0005-2795(75)90109-9] [Medline: 1180967]
38. Marzban C. The ROC curve and the area under it as performance measures. Weather Forecast 2004;19(6):1106-1114. [doi: 10.1175/825.1]
39. Mandrekar JN. Receiver operating characteristic curve in diagnostic test assessment. J Thorac Oncol 2010 Sep;5(9):1315-1316 [FREE Full text] [doi: 10.1097/JTO.0b013e3181ec173d] [Medline: 20736804]
40. Raschka S, Mirjalili V. Python Machine Learning. Birmingham, United Kingdom: Packt Publishing Ltd; 2017.
41. Longitudinal Assessment of Bariatric Surgery (LABS) Consortium, Flum DR, Belle SH, King WC, Wahed AS, Berk P, et al. Perioperative safety in the longitudinal assessment of bariatric surgery. N Engl J Med 2009 Jul 30;361(5):445-454 [FREE Full text] [doi: 10.1056/NEJMoa0901836] [Medline: 19641201]
42. Wilson AC, Roelofs R, Stern M, Srebro N, Recht B. arxiv preprints. 2017. The Marginal Value of Adaptive Gradient Methods in Machine Learning URL: https://arxiv.org/abs/1705.08292 [accessed 2020-03-04]
43. Marsland S. Machine Learning: An Algorithmic Perspective. Boca Raton, Florida: Chapman and Hall/CRC; 2011.
44. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015 Jan;61:85-117. [doi: 10.1016/j.neunet.2014.09.003] [Medline: 25462637]
45. Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw 2003;16(5-6):555-559. [doi: 10.1016/S0893-6080(03)00115-1] [Medline: 12850007]
46. Chollet F. Deep Learning with Python. Shelter Island, NY: Manning Publication Co; 2018.
47. Geubbels N, de Brauw LM, Acherman YI, van de Laar AW, Bruin SC. Risk stratification models: how well do they predict adverse outcomes in a large Dutch Bariatric Cohort? Obes Surg 2015 Dec;25(12):2290-2301. [doi: 10.1007/s11695-015-1699-2] [Medline: 25937046]
48. Greenstein AJ, Wahed AS, Adeniji A, Courcoulas AP, Dakin G, Flum DR, et al. Prevalence of adverse intraoperative events during obesity surgery and their sequelae. J Am Coll Surg 2012 Aug;215(2):271-7.e3 [FREE Full text] [doi: 10.1016/j.jamcollsurg.2012.03.008] [Medline: 22634116]
49. Thorell A, MacCormick AD, Awad S, Reynolds N, Roulin D, Demartines N, et al. Guidelines for perioperative care in bariatric surgery: enhanced recovery after surgery (ERAS) society recommendations. World J Surg 2016 Sep;40(9):2065-2083. [doi: 10.1007/s00268-016-3492-3] [Medline: 26943657]
50. Krishnamoorthy M, Suresh S, Alagappan S, Ahamed BB. Deep learning techniques and optimization strategies in big data analytics: automated transfer learning of convolutional neural networks using Enas algorithm. In: Thomas JJ, Karagoz P, Ahamed BB, Vasant P, editors. Deep Learning Techniques and Optimization Strategies in Big Data Analytics. Hershey, Pennsylvania: IGI Global; 2020:142-153.
51. Sünderhauf N, Brock O, Scheirer W, Hadsell R, Fox D, Leitner J, et al. The limits and potentials of deep learning for robotics. Int J Robot Res 2018;37(4-5):405-420. [doi: 10.1177/0278364918770733]

Abbreviations

1D: one-dimensional
Adam: adaptive moment estimation
AUC: area under curve
CAP: credit assignment path
CNN: convolutional neural network
DLNN: deep learning neural network
HbA1c: hemoglobin A1c
LSTM: long short-term memory
MCC: Matthews correlation coefficient
ML: machine learning
MLP: multilayer perceptron
relu: rectified linear unit
RNN: recurrent neural network
ROC: receiver operating characteristic
SMOTE: synthetic minority oversampling technique
SOReg: the Scandinavian Obesity Surgery Registry
WC: waist circumference

Edited by G Eysenbach; submitted 25.08.19; peer-reviewed by G Qin, E Kristiansson, L Zhang; comments to author 16.12.19; revised version received 07.01.20; accepted 07.02.20; published 08.05.20.

Please cite as:
Cao Y, Montgomery S, Ottosson J, Näslund E, Stenberg E
Deep Learning Neural Networks to Predict Serious Complications After Bariatric Surgery: Analysis of Scandinavian Obesity Surgery Registry Data
JMIR Med Inform 2020;8(5):e15992
URL: https://medinform.jmir.org/2020/5/e15992
doi:10.2196/15992
PMID:32383681

Original Paper

Prediction of Preeclampsia and Intrauterine Growth Restriction: Development of Machine Learning Models on a Prospective Cohort

Herdiantri Sufriyana[1,2], MD, MSc; Yu-Wei Wu[1,3], PhD; Emily Chia-Yu Su[1,3,4], PhD

[1]Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

[2]Department of Medical Physiology, College of Medicine, University of Nahdlatul Ulama Surabaya, Surabaya, Indonesia

[3]Clinical Big Data Research Center, Taipei Medical University Hospital, Taipei, Taiwan

[4]Research Center for Artificial Intelligence in Medicine, Taipei Medical University, Taipei, Taiwan

Corresponding Author:
Emily Chia-Yu Su, PhD
Graduate Institute of Biomedical Informatics
College of Medical Science and Technology
Taipei Medical University
No. 250 Wu-Xing Street
Taipei, 11031
Taiwan
Phone: 886 2 66382736 ext 1515
Email: emilysu@tmu.edu.tw

Abstract

Background: Preeclampsia and intrauterine growth restriction are placental dysfunction–related disorders (PDDs) that require a referral decision be made within a certain time period. An appropriate prediction model should be developed for these diseases. However, previous models did not demonstrate robust performances and/or they were developed from datasets with highly imbalanced classes.

Objective: In this study, we developed a predictive model of PDDs by machine learning that uses features at 24-37 weeks' gestation, including maternal characteristics, uterine artery (UtA) Doppler measures, soluble fms-like tyrosine kinase receptor-1 (sFlt-1), and placental growth factor (PlGF).

Methods: A public dataset was taken from a prospective cohort study that included pregnant women with PDDs (66/95, 69%) and a control group (29/95, 31%). Preliminary selection of features was based on a statistical analysis using SAS 9.4 (SAS Institute). We used Weka (Waikato Environment for Knowledge Analysis) 3.8.3 (The University of Waikato, Hamilton, NZ) to automatically select the best model using its optimization algorithm. We also manually selected the best of 23 white-box models. Models, including those from recent studies, were also compared by interval estimation of evaluation metrics. We used the Matthew correlation coefficient (MCC) as the main metric. It is not overoptimistic to evaluate the performance of a prediction model developed from a dataset with a class imbalance. Repeated 10-fold cross-validation was applied.

Results: The classification via regression model was chosen as the best model. Our model had a robust MCC (.93, 95% CI .87-1.00, vs .64, 95% CI .57-.71) and specificity (100%, 95% CI 100-100, vs 90%, 95% CI 90-90) compared to each metric of the best models from recent studies. The sensitivity of this model was not inferior (95%, 95% CI 91-100, vs 100%, 95% CI 92-100). The area under the receiver operating characteristic curve was also competitive (0.970, 95% CI 0.966-0.974, vs 0.987, 95% CI 0.980-0.994). Features in the best model were maternal weight, BMI, pulsatility index of the UtA, sFlt-1, and PlGF. The most important feature was the sFlt-1/PlGF ratio. This model used an M5P algorithm consisting of a decision tree and four linear models with different thresholds. Our study was also better than the best ones among recent studies in terms of the class balance and the size of the case class (66/95, 69%, vs 27/239, 11.3%).

Conclusions: Our model had a robust predictive performance. It was also developed to deal with the problem of a class imbalance. In the context of clinical management, this model may improve maternal mortality and neonatal morbidity and reduce health care costs.

(JMIR Med Inform 2020;8(5):e15411) doi:10.2196/15411

KEYWORDS

preeclampsia; intrauterine growth restriction; machine learning; uterine artery Doppler; sFlt-1/PlGF ratio

Introduction

Preeclampsia and intrauterine growth restriction (IUGR) are called placental dysfunction–related disorders (PDDs). These diseases have similar pathogeneses, biomarkers, and referral consequences [1,2]. However, they have different phenotypes and various correlations among biomarkers [3]. Subtypes of preeclampsia demonstrate heterogeneous gene expressions, yet a multiomics approach delineated no serological biomarkers [4]. These situations may cause difficulties in developing a robust prediction model for these diseases.

Preeclampsia prevalence ranges from 3% to 5% worldwide as a common disease contributing to maternal mortality [5]. The fetus of a pregnant woman with or without preeclampsia may undergo IUGR, which is associated with neonatal morbidity [6,7]. In spite of difficulties in distinguishing between these two diseases, both of them have similar consequences. They require referral to a hospital accompanied by advanced maternal and neonatal care within a certain time period [8]. Being able to predict PDDs would greatly support clinicians in making referral decisions, which should eventually improve both maternal and neonatal outcomes.

Compared to the traditional first-trimester screening, a prediction model is more reliable for women in several countries if it uses predictors in the second or third trimester. In those countries, women have low numbers of first visits in the first trimester [9]. Meanwhile, models for predicting PDDs have been developed mostly for preeclampsia at 11-13 weeks' gestation. This period is considered the best time window for its prediction and the most effective prevention method [10,11]. Therefore, if using only the first-trimester prediction, pregnant women in those countries lose the chance to undergo early screening of preeclampsia. Although prevention is still not available after the first trimester, the second- or third-trimester prediction will still impart benefits in the context of clinical management [12]. Decision on early delivery, including by cesarean section, was recommended in the cases of deteriorated maternal or fetal condition [13]. Pregnant women who are more likely to develop preeclampsia can achieve benefit by reaching out to hospitals with advanced maternal care within a certain time period if this condition was well predicted. This benefit is still achieved, although risk of preeclampsia is lately identified at the third trimester, particularly before term (ie, <37 weeks' gestation), in which early delivery will increase prematurity. Even though the babies were delivered at term from pregnant women who have developed IUGR, they still need advanced neonatal care. It is because low birth weight and in-hospital deaths were found to be more prevalent in those babies compared to those delivered from pregnant women without IUGR [14,15]. Nonetheless, previous models did not demonstrate robust predictive performances using features in any trimester and/or they were developed from datasets with highly imbalanced classes [16-27].

Predictive modeling using conventional statistical methods may be difficult for preeclampsia, since there are various correlations among its predictors [3]. As this disease has heterogeneous gene expressions, another possible difficulty is the noisy class of outcomes [4]. Machine learning methods are capable of dealing with such problems [28]. In addition, a common problem with preeclampsia and/or IUGR is a class imbalance, as models were shown to develop overoptimistic predictions [29]. This study attempted to develop a prediction method for PDDs by machine learning that uses features at 24-37 weeks' gestation, including maternal characteristics, uterine artery (UtA) Doppler measures, soluble fms-like tyrosine kinase receptor-1 (sFlt-1), and placental growth factor (PlGF).

Methods

Study Design

We developed a machine learning model and report it based on Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research [30]. Our study utilized a public dataset from a prospective cohort study based on STROBE (STrengthening the Reporting of OBservational studies in Epidemiology) guidelines [3]. We developed this model to predict a prognosis of pregnancy outcomes. The prediction model should solve a classification task between a control group and a cohort with a PDD, either preeclampsia or IUGR. A referral decision to a hospital with advanced care is a consequence related to an under- or overprediction of these diseases. Eventually, underprediction may increase maternal mortality and neonatal morbidity, while overprediction may increase health care costs as burdens to either patients or health insurance companies. We intended to avoid both of these scenarios. This goal can be considered to have been achieved if the prediction model demonstrates a higher Matthew correlation coefficient (MCC) than those of recent studies. The range of MCCs is from –1 (worst) to 1 (best). This metric can imply trade-off between underprediction (ie, lower sensitivity and higher specificity) and overprediction (ie, higher sensitivity and lower specificity). This trade-off is commonly evaluated by area under the receiver operating characteristic (ROC) curve (AUC) and accuracy. However, these metrics cannot fairly imply predictive performance in datasets with imbalanced classes [29], like preeclampsia and IUGR. For example, in a low-prevalence event (ie, 10/100, 10%), the predictive performances are still high in terms of sensitivity (ie, 9/10, 90%) and specificity (ie, 81/90, 90%) as parts of AUC. The accuracy (ie, 90/100, 90%) is also still high, but the MCC is not (ie, .62).

Data Source

The dataset used in this study is a public dataset in the Mendeley Data repository [31]. This dataset belongs to a study conducted at the University Medical Centre Ljubljana, Slovenia [3]. It was approved by the Republic of Slovenia National Medical Ethics Committee (No. 104/04/12). The original study collected data from September 2012 to January 2015. We downloaded this public dataset on March 11, 2019. Inclusion criteria were ≥24 weeks' gestation at the time of data collection and similar proportions of <34 or ≥34 weeks' gestation at delivery between

the PDD and control groups. For all women with a PDD, the time interval was 48 hours at maximum for the gestational age between data collection and delivery. Exclusion criteria were signs of prepregnancy hypertension, prepregnancy diabetes, hypertensive disorders during pregnancy, or gestational diabetes.

This dataset provides features (ie, predictors) consisting of maternal age (years), parity (nulliparous vs parous), maternal weight before pregnancy (kg), maternal height (m), BMI before pregnancy (kg/m^2), UtA Doppler measures, sFlt-1 (μg/L), PlGF (μg/L), and the sFlt-1/PlGF ratio. The UtA Doppler measures included the resistivity index (RI) of the UtA (RI-UtA), pulsatility index (PI) of the UtA (PI-UtA), and peak systolic velocity of the UtA (PSV-UtA). Each measure was taken for both the right and left UtAs. The average of both UtAs was calculated. In addition, the presence or absence of a bilateral notch was also included. The class (ie, outcome) consisted of 29 control subjects and 66 women with PDDs: 32 (48%) with both preeclampsia and IUGR, 12 (18%) with IUGR without preeclampsia, and 22 (33%) with preeclampsia without IUGR. Therefore, the ratio of positive (ie, PDD) to negative (ie, control) classes was 7:3. Detailed criteria for the ultrasound examination, blood sampling, and diagnosis of either preeclampsia or IUGR were previously described [3].

There were missing values in one subject for maternal weight, height, and BMI. However, the BMI classification was inferred from the report for that subject (ie, overweight) [3]. Considering the distribution of BMI before pregnancy, a feature was added by discretization (<25 kg/m^2 [underweight + normal] vs ≥25 kg/m^2 [overweight + obese]).

Feature Selection

We used SAS 9.4 (SAS Institute) to conduct preliminary statistical analyses. These intended to identify the relevancy of candidate features by their association with the class. The dataset with relevant features was initially used for comparison with machine learning models. To improve their predictive performance, we also used a built-in algorithm of feature selection in each model. Redundant features were removed using this algorithm. In addition, we compared the selected features with those from previous studies.

The association tests to identify the relevancy were conducted based on the data type. For categorical features, we used the Fisher exact test. For continuous features, the association test depended on the distributions in each class using the Kolmogorov-Smirnov normality test. Continuous features that were normally distributed in both classes (P≥.05) would be tested by an independent t test. If the variance was equal (P≥.05), we used the pooled method. Otherwise, we used the Satterthwaite method. For continuous features that were not normally distributed (P<.05), we used the Wilcoxon rank test. The features were significantly associated with the class if P<.05.

In addition to the association tests for scheme-independent feature selection or the filter method, we also conducted scheme-specific feature selection or the wrapper method using built-in algorithms in models as described in the Model Development section. Details on the algorithms of feature selection were meticulously described in Witten et al [32]. Complex model configurations, including to apply the algorithms, can be reproduced by entering the configuration code for each model (see Multimedia Appendix 1).

Model Development

We used Weka (Waikato Environment for Knowledge Analysis) 3.8.3 (The University of Waikato, Hamilton, NZ) to develop machine learning models. We chose this software because of its practical ability to compare multiple models at once. The predictive performance of a machine learning model can be affected by its configuration uncertainty. Considering this issue, we used an add-on package of Weka—Auto-Weka 2.6.1 (The University of British Columbia, Vancouver, CA). It automatically selects the best machine learning model [33]. Its algorithm optimizes the configuration of each model within a predefined time period based on a predefined evaluation metric. We defined the time period as 12 hours and the metric as the AUC. However, this package shows only the best model, which is not necessarily a white-box model that is easier for humans to understand. Therefore, we also manually selected the best among 23 white-box models. These models were in a default configuration. Details on configurations for automatically and manually selected models were described (see Multimedia Appendix 1).

Manual selection to decide the best white-box model consisted of three steps. In step 1, we analyzed models that had greater or equal predictive performance compared to the logistic regression as the baseline. We used a corrected resampled t test, which was modified from the conventional paired t test, as previously developed [32]. The modification was intended to correct the significance of the difference in each evaluation metric that increases because of an increasing k fold. To calculate the t statistic (see Equation 1), we calculated the difference ($\Delta\mu = \mu_{j1} - \mu_{j2}$) between the means of the metric from the first model (μ_{j1}) and those from the second model (μ_{j2}) trained by i_k and validated by j_k from k-fold validation as described in the Model Validation section. The variance was estimated by the average of the squared differences between the j_k metric for each model and the mean of both models: $\sigma_\delta^2 = (\sum [x_{j1} - \mu_j] + \sum [x_{j2} - \mu_j]) \div (2 \times n_j)$. The number of instances for the validation set was denoted as n_j.

$$t = \Delta\mu \div \sqrt{[\,(1 \div k + n_j \div n_i) \times \sigma_\delta^2\,]} \quad (1)$$

In step 2, after the list of compared models no longer shrank using the t test, we used interval estimates with a decimal point precision to further shrink it. In the last step, we chose the best model by focusing on its sensitivity, interpretability, and trade-off between sensitivity and specificity.

Since customization is not provided by Weka in some circumstances, we optimized the best model from the manual selection by determining a custom threshold. All subjects of the dataset were used to determine an initial threshold. We then adjusted it by cross-validation to pursue expected sensitivity and specificity that were empirically reliable for unobserved data. Only training subsets were used to adjust the threshold, while validation subsets were only used to evaluate the

predictive performances applying the predefined threshold. Details on the optimization procedure were also described (see Multimedia Appendix 1).

Model Validation

Internal validation was conducted by repeated 10-fold cross-validation. The dataset was randomized and split up into 10 subsets with similar class balances. We used nine subsets to train a model in each fold, while the remaining subsets were used to validate it. We repeated these folds for 100 iterations with different seeds of randomization sequences. Cross-validation estimates the predictive performance of external validation [34]. This method of internal validation also improves the reliability of the reported predictive performance [35].

In addition, we also validated the best model with a custom threshold. The validation set consisted of 10 new subsets (n=35) taken from the original dataset (N=95) by stratified random sampling in SAS 9.4. The class balance was similar among subsets. These subsets were used to customize a threshold in pursuit of expected sensitivity and specificity that were reliable in most of the subsets.

Evaluation Metrics

We applied multiple metrics to the model evaluation. These were calculated from a confusion matrix, which consists of true positives (TPs), true negatives (TNs), false negatives (FNs), and false positives (FPs). We calculated all of these metrics from recent studies because all of the metrics had not been reported. We inferred a confusion matrix from each study based on their sensitivity, specificity, and sample size of either positives (Ps) or negatives (Ns) (see Equations 2-5).

$$TP = P \times \text{Sensitivity}\ (\%) \quad (2)$$

$$FN = P - TP \quad (3)$$

$$TN = N \times \text{Specificity}\ (\%) \quad (4)$$

$$FP = N - TN \quad (5)$$

Point and interval estimates were used for comparison of each evaluation metric. Model selection was evaluated by the AUC, the area under the precision-recall curve (PRC), accuracy (see Equation 6), and sensitivity (see Equation 7). In addition, we evaluated the Akaike information criterion (AIC) to describe the trade-off between predictive performance and risk of overfitting relatively among models in the end of selection. The corrected AIC (AIC_C) was used, considering the small training set, as previously described [36,37]. The best model was also evaluated by a calibration plot. We then demonstrated an ROC curve of the well-calibrated model. Comparing our model to those from recent studies, we used the AUC, sensitivity, and specificity (see Equation 8), in addition to the selected metric, which was the MCC (see Equation 9), because those metrics were widely used. However, an evaluation by the MCC prevents misleading predictive performances, particularly in a model developed from datasets with imbalanced classes [29]. Class imbalance is a common situation in preeclampsia and IUGR studies. In this situation, the MCC can provide a fair evaluation when comparing prediction models in order to choose the one that shows optimal performances on both sensitivity and specificity.

$$\text{Accuracy}\ (\%) = (TP + TN) \div (TP + FN + TN + FP) \times 100\% \quad (6)$$

$$\text{Sensitivity}\ (\%) = (TP) \div (TP + FN) \times 100\% \quad (7)$$

$$\text{Specificity}\ (\%) = (TN) \div (TN + FP) \times 100\% \quad (8)$$

$$MCC = (TP \times TN - FN \times FP) \div \sqrt{(P \times [TP + FP] \times N \times [TN + FN])} \quad (9)$$

Results

Selected Features

Several features were selected based on a preliminary statistical analysis (see Table 1). Selected maternal characteristics were maternal weight before pregnancy, BMI values (kg/m^2), and BMI categories ($<25\ kg/m^2$ vs $\geq 25\ kg/m^2$). Other features included three measures of the RI-UtA, three measures of the PI-UtA, the presence or absence of a bilateral notch, sFlt-1, PlGF, and the sFlt-1/PlGF ratio. The best model was automatically selected by a correlation-based feature selection of subset evaluation. It was combined with a backward greedy stepwise search algorithm.

The selected features were extracted from mostly similar measures in recent studies (see Table 2). These were maternal characteristics, PI-UtA, sFlt-1, and PlGF, but not the bilateral notch. The sFlt-1/PlGF ratio turned out to be the most important feature in the best model (see Figure 1) as previously described [1,38,39].

However, the best model by manual selection was the right PI-UtA over the mean value. This choice is counterintuitive if the placental side is contralateral to the side on which the PI-UtA was measured. A previous study found that the PI-UtA was lower on the side ipsilateral to the placental side [40]. We then added the lowest value as a feature to provide an acceptable measure of the PI-UtA regardless of the placental laterality. We also demonstrated the proportion of the PI-UtA as the lowest value in either the right or left UtA (see Table 1). In this study, most of the lowest PI-UtA values were found in the right UtA (66/95, 69%).

Table 1. Descriptive and comparative analyses.

Feature	Class		*P* value
	Control (n=29)	PDDs[a] (n=66)	
Maternal characteristics			
Maternal age (years), mean (95% CI)[b]	31.2 (30.9-31.5)	32.6 (32.4-32.7)	.23[c]
Parity, n (%)[d]			**.10[e]**
Nulliparous	15 (52)	47 (71)	
Parous	14 (48)	19 (29)	
Maternal weight (kg), median (IQR)[f]	58.0 (55.0-65.0)	68.0 (60.0-76.0)	.001[g,h]
Maternal height (m), mean (95% CI)	1.66 (1.658-1.666)	1.65 (1.651-1.655)	.51[c]
BMI (kg/m^2), median (IQR)	21.6 (19.9-22.5)	24.4 (23.0-28.2)	<.001[g,h,i]
BMI, n (%)			**.01[e,g]**
<25 kg/m^2	24 (83)	36 (55)	
≥25 kg/m^2	5 (17)	30 (45)	
Uterine artery (UtA) Doppler measures, median (IQR)			
Right resistivity index (RI)-UtA	0.57 (0.49-0.61)	0.71 (0.63-0.78)	<.001[g,h]
Left RI-UtA	0.59 (0.53-0.64)	0.73 (0.61-0.78)	<.001[g,h]
Mean RI-UtA	0.57 (0.52-0.62)	0.71 (0.61-0.77)	<.001[g,h]
Right pulsatility index (PI)-UtA	0.66 (0.60-0.71)	1.24 (0.79-1.56)	<.001[g,h,i]
Left PI-UtA	0.70 (0.67-0.75)	1.33 (0.82-1.59)	<.001[g,h]
Mean PI-UtA	0.68 (0.63-0.71)	1.26 (0.86-1.57)	<.001[g,h,i]
Right peak systolic velocity (PSV)-UtA	58.30 (55.10-62.40)	59.25 (56.80-64.18)	.09[h]
Left PSV-UtA	60.20 (59.10-64.10)	60.05 (57.10-63.80)	.99[h]
Mean PSV-UtA	59.55 (58.25-61.40)	60.38 (57.54-64.06)	.31[h]
Bilateral notch, n (%)			**<.001[e,g,i]**
Nulliparous	0 (0)	47 (71)	
Parous	29 (100)	19 (29)	
Lowest PI-UtA, median (IQR)	0.65 (0.57-0.69)	1.16 (0.74-1.53)	<.001[g,h,j]
Laterality of lowest PI-UtA, n (%)			**.23[e]**
Right UtA	23 (79)	43 (65)	
Left UtA	6 (21)	23 (35)	
sFlt-1[k] and PlGF [l], median (IQR)			
sFlt-1 (µg/L)	3014 (1852-4116)	13,961 (8893-22,218)	<.001[g,h,i]
PlGF (µg/L)	626.9 (281.3-752.8)	68.4 (42.9-150.1)	<.001[g,h,i]
sFlt-1/PlGF ratio	4.7 (2.6-15.1)	230.1 (100.8-483.0)	<.001[g,h,i]

[a]PDD: placental dysfunction–related disorder.

[b]Mean and 95% CI were calculated for numerical values with a normal distribution.

[c]Independent *t* test.

[d]Numbers and column proportions (%) were calculated for categorical values.

[e]Fisher exact test.

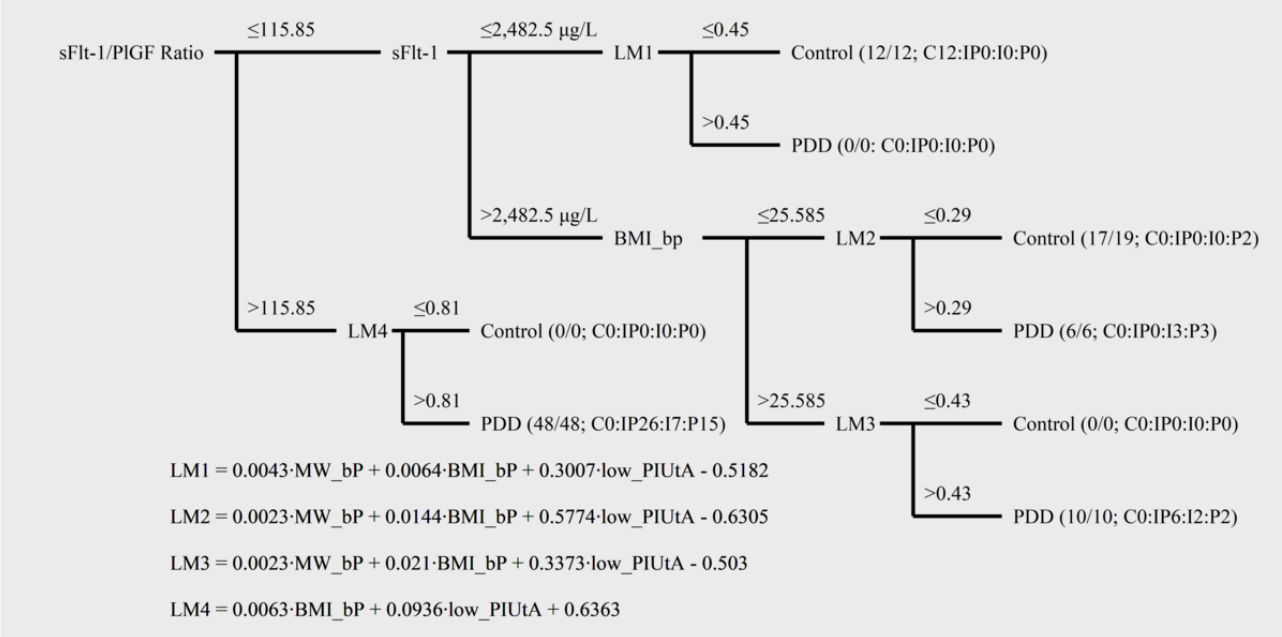
[f]Median and IQR were calculated for numerical values without a normal distribution.
[g]Statistically significant (alpha=.05).
[h]Wilcoxon rank test.
[i]Selected feature for the best model from automatic selection.
[j]Used for manual selection only.
[k]sFlt-1: soluble fms-like tyrosine kinase receptor-1.
[l]PlGF: placental growth factor.

Table 2. Features used by the models in this study compared to those from previous studies.[a]

Source	Gestational age at prediction	Features, n (for maternal characteristics) or + (used by the model) or – (not used by the model)						
		Maternal characteristics	MAP[b]	PI-UtA[c]	Bilateral notch	sFlt-1[d]	PlGF[e]	PAPP-A[f]
This study								
CVR[g]1 (right PI-UtA)	24-37 weeks	2	–	+	–	+	+	–
CVR2 (mean PI-UtA)	24-37 weeks	2	–	+	+	+	+	–
CVR3 (lowest PI-UtA)	24-37 weeks	2	–	+	–	+	+	–
158-tree random forest	24-37 weeks	1	–	+	+	+	+	–
Previous studies								
Wright A et al (2019) [26]	11-13 weeks	10	–	+	–	+	+	–
Wright D et al (2019) [27]	11-13 weeks	11	+	+	–	+	+	–
Tan MY et al (2018) [25]	11-13 weeks	11	+	+	–	–	+	–
Sonek J et al (2018) [24]	11-13 weeks	10	–	+	–	–	+	+
Perales A et al (2017) [23]	27-28 weeks	3	+	–	–	+	+	–
Nuriyeva G et al (2017) [22]	11-13 weeks	N/A[h]	–	+	–	–	+	+
O'Gorman N et al (2017) [21]	11-13 weeks	11	+	+	–	–	+	+
Gallo DM et al (2016) [18]	19-24 weeks	11	+	+	–	+	+	–
Tsiakkas A et al (2016) [19]	30-34 weeks	11	–	–	–	+	+	–
Andrietti S et al (2016) [20]	35-37 weeks	11	+	+	–	+	+	–
O'Gorman N et al (2016) [17]	11-13 weeks	10	+	+	–	–	+	+
Wright D et al (2015) [16]	11-13 weeks	11	–	–	–	–	–	–

[a]Models that showed the best sensitivity and an acceptable specificity in each study.
[b]MAP: mean arterial pressure.
[c]PI-UtA: pulsatility index of the uterine artery.
[d]sFlt-1: soluble fms-like tyrosine kinase receptor-1.
[e]PlGF: placental growth factor.
[f]PAPP-A: pregnancy-associated plasma protein-A.
[g]CVR: classification via regression.
[h]N/A: not applicable.

Figure 1. Characteristics of the classification via regression model using the lowest pulsatility index of the uterine artery (PI-UtA). Fractions in leaf nodes consist of true predicted numbers (numerators) and all predicted ones (denominators). A ratio of true predicted numbers is shown for control (C), both intrauterine growth restriction (IUGR) and preeclampsia (IP), IUGR only (I), and preeclampsia only (P). BMI_bP: body mass index before pregnancy (kg/m^2); LM: linear model; low_PIUtA: the lowest pulsatility index of the uterine artery; MW_bP: maternal weight before pregnancy (kg); PDD: placental dysfunction–related disorder; PlGF: placental growth factor; sFlt: soluble fms-like tyrosine kinase receptor.



Selected Machine Learning Models

We focused on the sensitivity to ensure minimum miss rates, which should improve maternal and neonatal outcomes. This resulted in the seven best machine learning models as shown in Table 3. The best model was the random forest from automatic selection; however, it is not a white-box model. We then also manually selected the best white-box model.

Classification via regression (CVR) classifies an outcome based on an M5P regression algorithm. It combines a pruned decision tree with smoothed linear models. There is also a built-in algorithm in CVR for selecting important features. A feature at the root node of the decision tree is the most important. Each leaf node has different linear models (LMs), which can be set to use different thresholds [32]. Optimization of this model was conducted by determining these thresholds (see Multimedia Appendix 1).

We developed CVR using only the mean values of UtA Doppler measures, in addition to this model using the right PI-UtA. We also developed CVR using the lowest PI-UtA value without other UtA Doppler measures. In the end, the model using the lowest PI-UtA value (see Figure 1) was the best, followed by that using either the right or mean PI-UtA (see Multimedia Appendices 2 and 3). We provided an interactive interface for readers to apply the model using the lowest PI-UtA value (see Multimedia Appendix 4).

We demonstrated characteristics of the best CVR using selected features from all subjects of the dataset (see Figure 1). LM1, LM3, and LM4 perfectly classified outcomes. However, a subpopulation of subjects was misclassified as the control instead of as having isolated preeclampsia. It consisted of subjects with sFlt-1/PlGF of ≤115.85, sFlt-1 of >2482.5 µg/L, and a BMI of ≤25.585 kg/m^2.

Calibration plots are shown for CVR models using different types of PI-UtA (see Figure 2). Positive samples gathered higher values of both predicted and true probabilities from all of the CVR models. Then, classification biases were higher on positive samples from these models. However, all of the biases remained low because the root mean square error (RMSE) was only 0.076 at the maximum upper bound of the subsets, particularly from CVR using the mean PI-UtA. Therefore, these models were well calibrated. They also indicated robust positive predictive values (PPVs) or information retrieval (IR) precision.

ROC curves are also shown for the CVR models (see Figure 3). C-statistics of 10 subsets are represented by an AUC that is shown for each CVR model. An average sensitivity was calculated for each distinct value of FP rates in order to measure the AUCs. The greatest AUC was for the CVR model that used the lowest PI-UtA (see Table 4). It significantly differs from that of the model using the right or mean PI-UtA value. Applying different thresholds for each LM, each CVR model has an acceptable trade-off between sensitivity and specificity without compromising its MCC.

Table 3. The seven best machine learning models.

Model	Performance metrics and rank				
	Area under the ROC[a] curve	Area under the PRC[b]	Accuracy (%)	Δ_i AIC_C[c]	Sensitivity (%)
Automatic selection: random forest	0.976 (1)	0.958 (1)	92.6 (1)	0 (1)	90.7 (1)
Manual selection					
CVR[d]	0.954 (5)	0.922 (3)	90.6 (4)	15 (4)	89.7 (2)
Naïve Bayes	0.960 (2)	0.928 (2)	90.2 (5)	25 (5)	89.0 (3)
Simple logistic	0.958 (3)	0.921 (4)	90.9 (2)	6 (2)	88.2 (4)
Logistic model tree	0.957 (4)	0.920 (5)	90.8 (3)	7 (3)	88.0 (5)
Multi-class classifier	0.932 (6)	0.868 (6)	89.9 (6)	30 (6)	86.8 (6)
Logistic regression	0.932 (7)	0.868 (7)	89.9 (7)	30 (7)	86.8 (7)

[a]ROC: receiver operating characteristic.

[b]PRC: precision-recall curve.

[c]AIC_C: corrected Akaike's information criterion (Δ_i AIC_C = AIC_{Ci} – $AIC_{C\ min}$).

[d]CVR: classification via regression.

Figure 2. Calibration plots of classification via regression (CVR) models using the lowest, right, and mean pulsatility index of the uterine artery (PI-UtA). Each point demonstrates a validation subset taken from repeated 10-fold cross-validation. Colors denote subsets from stratified random sampling. RMSE: root mean square error.

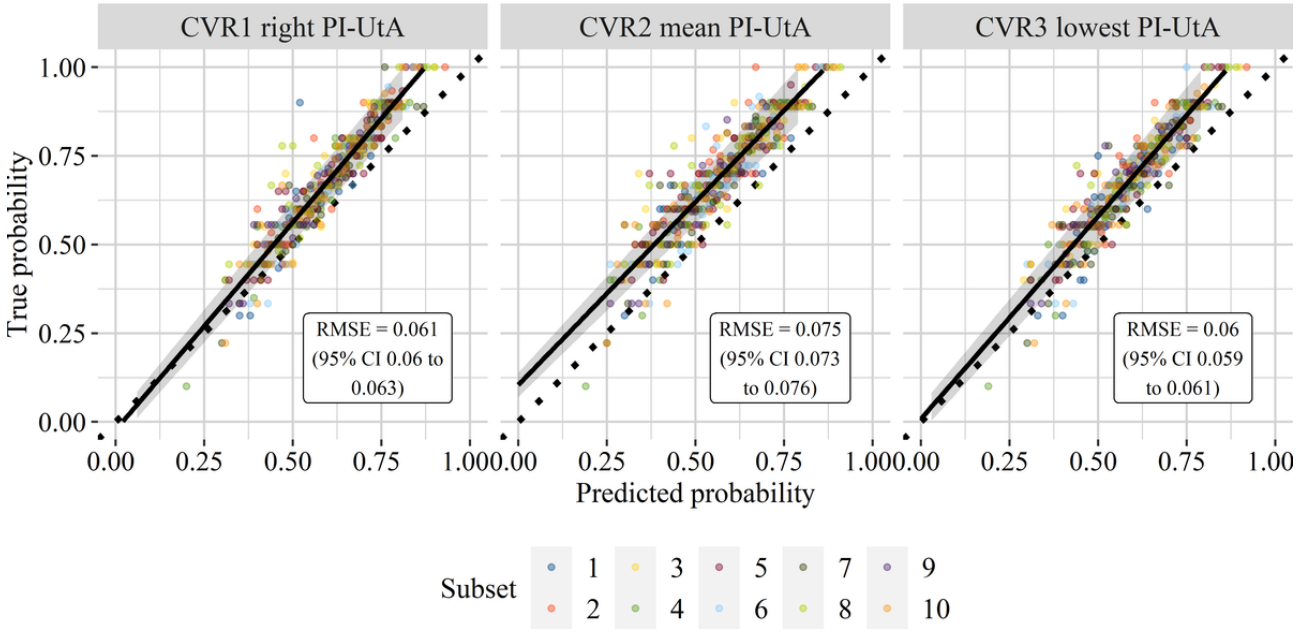


Table 4. Predictive performances shown by models in this study compared to those from recent studies.[a]

Source	Predictive performance[b]		
	AUC[c]	Sensitivity, %	Specificity, %
This study			
CVR[d]1 (right PI-UtA[e])	0.906 (0.896-0.916)	91 (85-96)	97 (90-100)
CVR2 (mean PI-UtA)	0.926 (0.919-0.933)	95 (91-100)	100 (100-100)
CVR3 (lowest PI-UtA)	0.970 (0.966-0.974)	95 (91-100)	100 (100-100)
158-tree random forest	0.976 (0.967-0.985)	91 (87-94)	93 (92-95)
Recent studies			
Wright A et al (2019) [26]	N/A[f,g]	85 (72-94)	90 (90-90)
Wright D et al (2019) [27]	0.970 (0.950-0.990)	93 (76-99)	90[h]
Tan MY et al (2018) [25]	N/A[g]	90 (80-96)	90[h]
Sonek J et al (2018) [24]	N/A[g]	85[i]	95[i]
Perales A et al (2017) [23]	0.930[i]	81[i]	95[i]
Nuriyeva G et al (2017) [22]	0.888[i]	76[i]	90[i]
O'Gorman N et al (2017) [21]	0.987[i]	100 (80-100)	90[h]
Gallo DM et al (2016) [18]	0.930 (0.892-0.968)	85 (74-93)	90[h]
Tsiakkas A et al (2016) [19]	0.987 (0.980-0.994)	100 (92-100)	90[h]
Andrietti S et al (2016) [20]	0.938 (0.917-0.959)	82 (70-91)	90[h]
O'Gorman N et al (2016) [17]	0.907[i]	89 (79-96)	90[h]
Wright D et al (2015) [16]	0.811[i]	67 (59-74)	90[h]

[a]Models that showed the best sensitivity and an acceptable specificity in each study.

[b]Point and interval estimates.

[c]AUC: area under the receiver operating characteristic (ROC) curve.

[d]CVR: classification via regression.

[e]PI-UtA: pulsatility index of the uterine artery.

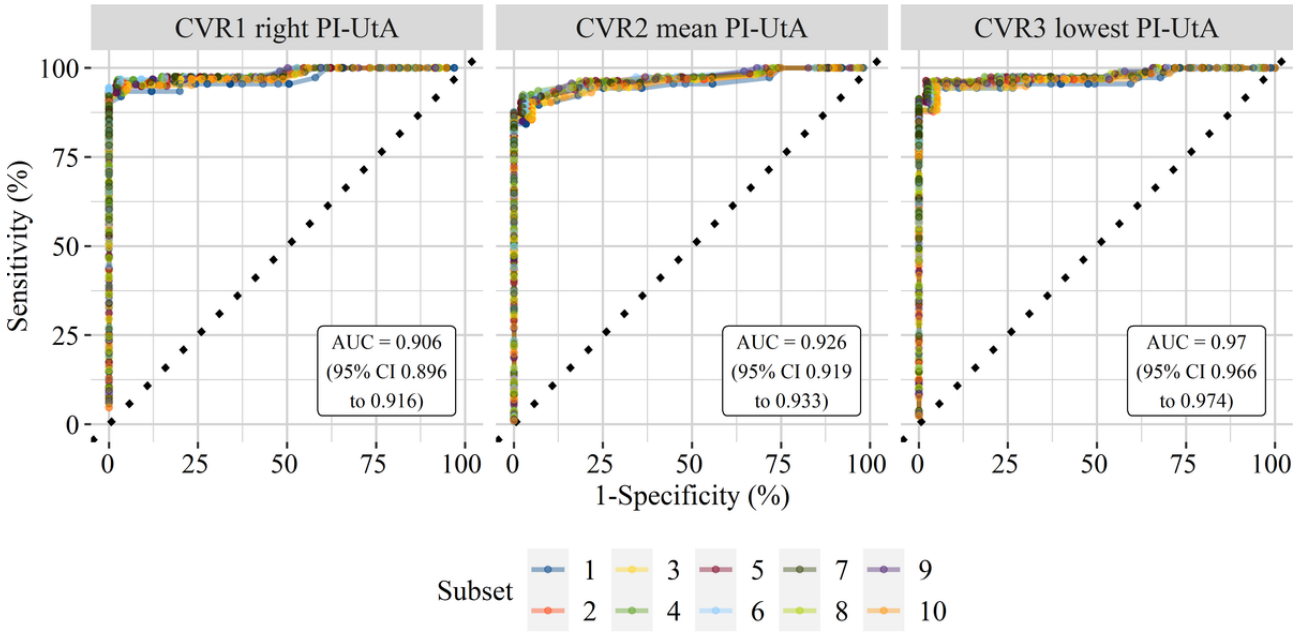
[f]N/A: not applicable because it was not available.

[g]This study showed an ROC curve without an AUC statement.

[h]Fixed specificity in order to define sensitivity.

[i]This study did not report an interval estimate.

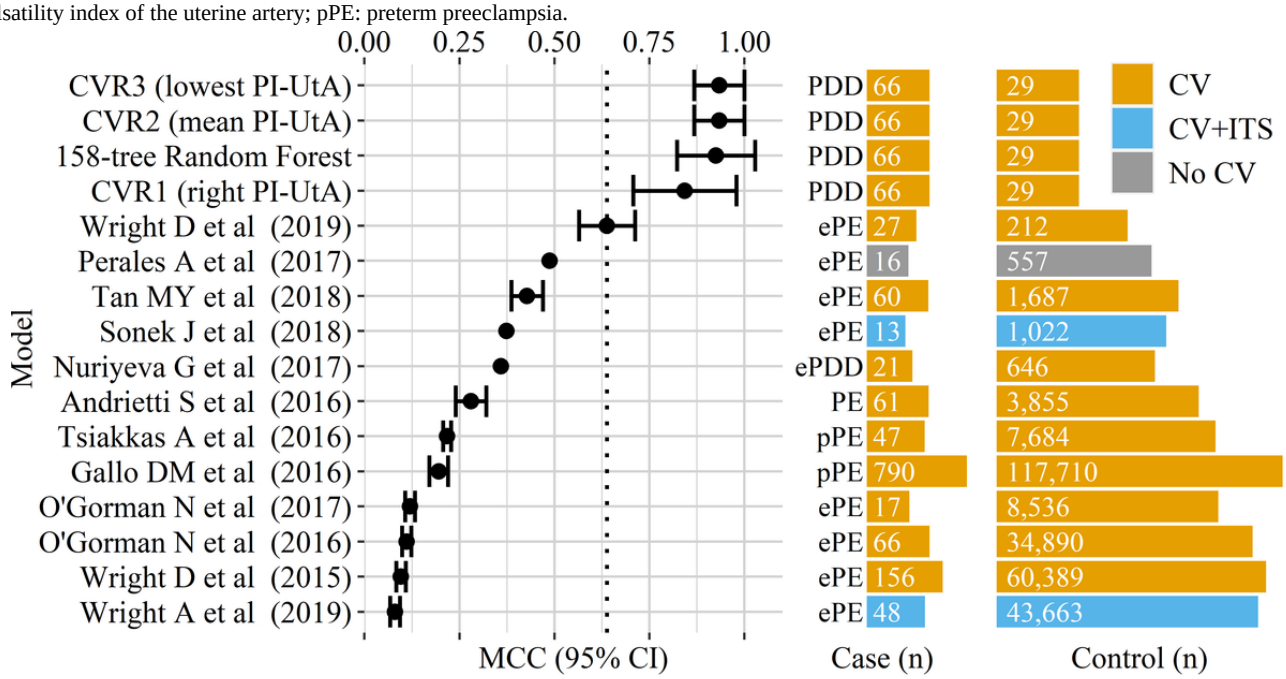
Figure 3. Receiver operating characteristic (ROC) curves of classification via regression (CVR) models using the lowest, right, and mean pulsatility index of the uterine artery (PI-UtA). Each ROC curve demonstrates a validation subset taken from repeated 10-fold cross-validation. Colors denote subsets from stratified random sampling. AUC: area under the receiver operating characteristic curve.



Comparison of Predictive Performances

The CVR model with the lowest PI-UtA value was found to achieve the most robust predictive performance (see Figure 4 and Table 4), as determined by the MCC (.93, 95% CI .87-1.00). The MCC of this model showed no difference compared to that of either the best model from automatic selection (.93, 95% CI .82-1.00) or the CVR model with the mean PI-UtA value (.93, 95% CI .87-1.00). However, the MCC of this CVR model was higher than those from the models with the right PI-UtA value (.84, 95% CI .71-.98). The predictive performance in this study was assessed by cross-validation without an independent test set, similar to most of the recent studies. However, we developed our models from a dataset with a class balance that was better than those of recent studies. The MCCs of our models were also higher than those of recent studies (see Figure 4 and Multimedia Appendix 3). Compared to random forest with the best AIC (see Table 3), the CVR models with the lowest, right, and mean PI-UtA showed AIC values of 13, 15, and 17, respectively.

Figure 4. The Matthew correlation coefficient (MCC) and class balance. Control samples did not include other subtypes of either hypertension in pregnancy or placental dysfunction–related disorders (PDDs). Colors denote validation methods. Several studies did not report interval estimates and/or cross-validation (CV). To improve visualization, the scales for either case or control sample sizes were individually log-transformed. CVR: classification via regression; ePDD: early placental dysfunction–related disorder; ePE: early preeclampsia; ITS: independent test set; PE: preeclampsia; PI-UtA: pulsatility index of the uterine artery; pPE: preterm preeclampsia.



Comparison of predictive performances was also described using other evaluation metrics that are commonly used (see Table 4). There was significant difference in the AUC between the CVR models that used the lowest and other PI-UtA values. Meanwhile, the CVR model with the lowest PI-UtA value was not significantly different compared to the automatically selected 158-tree random forest. From recent studies, Wright et al [27] and Tsiakkas et al [19] showed models with more competitive areas under the ROC than those of our models. However, our models show sensitivities and specificities that are not inferior compared to those from recent studies. In addition, our models were developed by a dataset with a better class balance, whose case class size was 69% (66/95), compared to the most balanced dataset from Wright et al [27], whose case size was 11.3% (27/239) (see Figure 4).

Discussion

Principal Findings

The best model in this study was a CVR one that used the lowest PI-UtA values. It was an acceptable model, because the lowest PI-UtA value was reliably found ipsilateral to the placental side [40]. This model demonstrated higher MCCs and PPVs, but not sensitivity or AUC, compared to those from previous studies (see Figure 4, Table 4, and Multimedia Appendix 3). MCC was intended for achieving our goal to eventually avoid mortality and morbidity and unnecessary health care costs. This may result in improved maternal and neonatal outcomes. It also outperformed models from recent studies in terms of specificity. Compared to a model that had 90% specificity, this potentially reduces 10% of health care costs. Applying a predictive model that uses the sFlt-1/PlGF ratio, a previous study showed a similar reduction in health care costs [41]. Even without considering the health economics, the MCC is still practical to consider FPs along with other components of the confusion matrix, which reflect numbers of false referral decisions on predicted preeclampsia and IUGR. Making wrong decisions may harm pregnant women, especially in developing countries where a distant and dangerous journey must be taken by pregnant women to reach higher-level health care facilities. Therefore, a CVR model that used the lowest PI-UtA values was better in compromising between the mortality and morbidity and costs compared to the those of other models in either this study or previous studies.

Comparison With Prior Work

The selected features were consistent with those from previous studies. The preeclampsia risk was found to be higher in women with a prepregnancy BMI classified as overweight or obese compared to those classified as underweight or normal (with a cutoff of ≥24 kg/m^2) [42]. This disease was also associated with combinations of a bilateral notch, both RI-UtA and PI-UtA, and sFlt-1/PlGF measures in the second or third trimester [43,44]. However, these combinations were inconsistently associated with the IUGR with or without preeclampsia [45-47]. As to the UtA Doppler measures, no association was found between placental location and either preeclampsia or a low birth weight [48]. Using features corresponding to results from previous studies, an acceptable machine learning model can be developed. CVR belongs to a group of superior meta-classifiers for predicting malicious cyberattacks, but it was not the best as a bagging classifier [49]. In this study, the bagging classifier did not outperform CVR. The optimized CVR model was also better than the random forest from automatic selection. Surprisingly, this model was not outperformed by any state-of-the-art machine learning models. Those included both artificial neural networks and support vector machines. These models were also candidates for automatic selection in this study. One possible reason is because of a regression model used by CVR that divides the dataset into several subpopulations using a decision tree. In the field of medicine, this algorithm is widely known as a reliable and effective machine learning application [50].

Each leaf node in the decision tree has a different LM. It can capture different correlations among features in each subpopulation that is normally distributed [51]. Different thresholds for each LM may approach heterogeneity in PDDs, especially in preeclampsia. Thresholds or cutoffs also give more understanding as to how outcomes are predicted. Thus, this model has the interpretability that we intended to achieve.

In this study, the CVR models split subjects by an sFlt-1/PlGF ratio of 115.85. This cutoff was higher than 38 as previously described [38,39]. This is reasonable, because predicted outcomes in this study were not only preeclampsia but also IUGR. Birth weights showed no difference for babies from women with IUGR that were classified by 38 as a cutoff for the sFlt-1/PlGF ratio [47]. Therefore, a different cutoff for the sFlt-1/PlGF ratio is related to predicted outcomes in this study that differed from those of previous studies.

PIs were also selected by the CVR models of UtA Doppler measures. Unexpectedly, one of the CVR models in this study chose the right PI-UtA instead of the mean value, which is conventionally used [27,44,47]. This is counterintuitive because of placental laterality, although a previous study showed no difference between the right and left PI-UtA values (P=.20) [52]. However, the CVR model using the lowest value had a higher MCC than that using the right PI-UtA in this study. A previous model demonstrated a greater AUC when using the lowest PI-UtA instead of the mean or highest value [53]. This is also more acceptable, because the lowest PI-UtA value was shown to be ipsilateral to the placental location [40]. Thus, this measure is independent of placental laterality.

However, between the CVR model using the right PI-UtA and the one using the lowest value, we may also consider several similarities. These were shown by most of the evaluation metrics and characteristics. The similarities may be coincidental because most of the subjects had the lowest value on the right side of the UtA in this study (66/95, 69%; see Table 1). Most placentas were located on the right side (57.4%) compared to the middle (22.2%) and left side (20.4%) on the anterior uterine wall [54]. Interestingly, the sleeping position before becoming pregnant was mostly right lateral by pregnant women with a placenta on either the anterior, lateral, or fundal uterine wall (P=.001) [55].

In addition to the lowest and the right PIs, the CVR model using the mean PI-UtA value also demonstrated a competitive predictive performance. This model showed each LM using a combination of the mean PI-UtA and bilateral notch. Apparently,

both of them are a counterpart of the lowest or the right PI-UtA alone in each LM of other CVR models. The predictive value of the mean PI-UtA was found to be higher if the bilateral notch was present compared to when it was absent [43]. Nevertheless, this model demonstrated the highest RMSE compared to CVR models using the lowest or the right PI-UtA (see Figure 2). Therefore, the best model in this study was the CVR model that used the lowest PI-UtA.

The best model used 25.585 kg/m^2 as a cutoff for BMI in its decision tree. This is similar to the cutoff for BMI as a risk factor of preeclampsia [42]. As indicated by each LM in the best model, an effect on PDDs was partially contributed by the two maternal characteristics of maternal weight and the BMI. However, the risk of preeclampsia, as a subtype of PDD, was adjusted by multiple factors instead of only these anthropometrics [56]. Other maternal characteristics were not represented in the dataset we used. So, our models need further improvement using a dataset with more maternal characteristics.

None of the predictive models from 12 recent studies outperformed our models according to the MCC [16-27]. All of those studies used datasets with highly imbalanced classes that may have masked the misclassification of positive samples [29]. There are many aspects that may cause similar problems [3,4,28]. These include an outcome leakage that was encountered by some of those studies [18,20,23]. Mean arterial pressure (MAP) may easily infer the class because it is calculated from the same measures as for the diagnostic criteria of preeclampsia. This is true if MAP is taken in the second trimester, when it is used for predicting either early or preterm preeclampsia. This feature may also cause an outcome leakage if it is taken at 35-37 weeks' gestation, when it is used for predicting late preeclampsia. Outcome leakage causes the predictive performance to be overoptimistic [30].

Strengths

To the best our knowledge, this is the first study that used machine learning to predict preeclampsia and/or IUGR using features in the second or third trimester of pregnancy. Our models outperformed 12 recent studies according to the MCC. This study also used a dataset with a better class balance than those used by recent studies as well as the size of the case class. Predicting preeclampsia [26,27] and IUGR [47] used to be developed using conventional statistical modeling. A previous study developed a machine learning model (ie, multilayer perceptron) for predicting PDDs in the first trimester [22]. However, its PPV or IR precision was insufficient. Other studies developed a machine learning model to characterize gene expression of preeclampsia as mechanism studies instead of for prediction [4,57]. Yet, a machine learning model can both perform a robust prediction and reveal mechanisms of a disease.

Limitations

A pitfall should be considered when applying our models. They do not distinguish between preeclampsia and IUGR. These models should only be applied for a referral decision. This means whether a clinician should refer the pregnant women to a hospital with advanced maternal and neonatal care within a certain time period [8]. For pregnant women who will develop preeclampsia with or without IUGR before term, advanced maternal care will be needed for cesarean section. It is one of the possible modes for early delivery that was recommended at any time in deteriorated maternal or fetal condition [13]. Meanwhile, for pregnant women who will develop IUGR with or without preeclampsia, the advanced neonatal care will be needed for the babies. They were found having low birth weight and more in-hospital deaths, even among those who were delivered at term [14,15].

Other applications of our models exclude a decision of delivery before term. This decision should be made based on models that specifically predict severe cases of early-onset or preterm preeclampsia and IUGR. It is because a false decision on early delivery will bring unnecessary prematurity. Nonetheless, no prediction for isolated preeclampsia is needed for those at term since no prematurity will occur as a consequence of early delivery decision.

Controls in this study also did not include other subtypes of hypertension in pregnancy. They may be indistinguishable from PDDs, but there is no need for patient referral. There is a possibility that more FPs will occur in subjects who will develop other subtypes of this disease. Therefore, the clinical impact may be unnecessary patient referral to higher-level health care facilities.

We also need to conduct external validation to confirm predictive performance of our models. There is a possibility that these models overfit the dataset. This is still possible even though they were evaluated by sufficient cross-validation because of consideration of diverse phenotypes of preeclampsia, other subtypes of hypertension in pregnancy, and other PDDs.

Conclusions

CVR is a machine learning model that has robust predictive performance in classifying PDDs versus a control group. This model differentiates PDDs from a control that has no other subtypes of hypertension in pregnancy. Using features in the second or third trimester, this model may be reliable for countries with low numbers of first visits in the first trimester, but further investigations are needed. Although the best preventive method for preeclampsia is not in the second or third trimester, this model can still be beneficial in the context of clinical management.

Acknowledgments

Tanja Premru-Srsen from the Department of Perinatology, Division of Obstetrics and Gynecology, University Medical Centre Ljubljana, Slovenia, provided the dataset in this study through the Mendeley Data repository. She also provided a reprint of the full text that initially reported this dataset. This study was funded by the Ministry of Science and Technology (MOST) in Taiwan (grant numbers MOST107-2221-E-038-016 and MOST108-2221-E-038-018) and the Higher Education Sprout Project by the

Ministry of Education (MOE) in Taiwan (grant numbers DP2-107-21121-01-A-01 and DP2-108-21121-01-A-01-04), with funding awarded to Emily Chia-Yu Su. The sponsors had no role in the research design or contents of the manuscript for publication.

Conflicts of Interest

None declared.

Multimedia Appendix 1
Automatic and manual model selection.
[DOCX File , 373 KB - medinform_v8i5e15411_app1.docx]

Multimedia Appendix 2
Characteristics of classification via regression (CVR) models with the right and mean of the pulsatility index of the uterine artery (PI-UtA).
[DOCX File , 282 KB - medinform_v8i5e15411_app2.docx]

Multimedia Appendix 3
Evaluation metrics and validation method for comparison with recent studies.
[DOCX File , 24 KB - medinform_v8i5e15411_app3.docx]

Multimedia Appendix 4
Interactive model.
[PDF File (Adobe PDF File), 125 KB - medinform_v8i5e15411_app4.pdf]

References

1. Kwiatkowski S, Dołegowska B, Kwiatkowska E, Rzepka R, Marczuk N, Loj B, et al. Maternal endothelial damage as a disorder shared by early preeclampsia, late preeclampsia and intrauterine growth restriction. J Perinat Med 2017 Oct 26;45(7):793-802. [doi: 10.1515/jpm-2016-0178] [Medline: 27865093]
2. Reijnders IF, Mulders AGMGJ, Koster MPH. Placental development and function in women with a history of placenta-related complications: A systematic review. Acta Obstet Gynecol Scand 2018 Mar;97(3):248-257. [doi: 10.1111/aogs.13259] [Medline: 29125627]
3. Fabjan-Vodusek V, Kumer K, Osredkar J, Verdenik I, Gersak K, Premru-Srsen T. Correlation between uterine artery Doppler and the sFlt-1/PlGF ratio in different phenotypes of placental dysfunction. Hypertens Pregnancy 2019 Feb;38(1):32-40. [doi: 10.1080/10641955.2018.1550579] [Medline: 30485134]
4. Nair TM. Statistical and artificial neural network-based analysis to understand complexity and heterogeneity in preeclampsia. Comput Biol Chem 2018 Aug;75:222-230. [doi: 10.1016/j.compbiolchem.2018.05.011] [Medline: 29859381]
5. Abalos E, Cuesta C, Grosso AL, Chou D, Say L. Global and regional estimates of preeclampsia and eclampsia: A systematic review. Eur J Obstet Gynecol Reprod Biol 2013 Sep;170(1):1-7. [doi: 10.1016/j.ejogrb.2013.05.005] [Medline: 23746796]
6. Class QA, Rickert ME, Lichtenstein P, D'Onofrio BM. Birth weight, physical morbidity, and mortality: A population-based sibling-comparison study. Am J Epidemiol 2014 Mar 01;179(5):550-558 [FREE Full text] [doi: 10.1093/aje/kwt304] [Medline: 24355331]
7. Nardozza LMM, Caetano ACR, Zamarian ACP, Mazzola JB, Silva CP, Marçal VMG, et al. Fetal growth restriction: Current knowledge. Arch Gynecol Obstet 2017 May;295(5):1061-1077. [doi: 10.1007/s00404-017-4341-9] [Medline: 28285426]
8. von Dadelszen P, Payne B, Li J, Ansermino JM, Broughton Pipkin F, Côté AM, PIERS Study Group. Prediction of adverse maternal outcomes in pre-eclampsia: Development and validation of the fullPIERS model. Lancet 2011 Jan 15;377(9761):219-227. [doi: 10.1016/S0140-6736(10)61351-7] [Medline: 21185591]
9. Moller A, Petzold M, Chou D, Say L. Early antenatal care visit: A systematic analysis of regional and global levels and trends of coverage from 1990 to 2013. Lancet Glob Health 2017 Oct;5(10):e977-e983 [FREE Full text] [doi: 10.1016/S2214-109X(17)30325-X] [Medline: 28911763]
10. Park F, Russo K, Williams P, Pelosi M, Puddephatt R, Walter M, et al. Prediction and prevention of early-onset pre-eclampsia: Impact of aspirin after first-trimester screening. Ultrasound Obstet Gynecol 2015 Oct;46(4):419-423 [FREE Full text] [doi: 10.1002/uog.14819] [Medline: 25678383]
11. Roberge S, Nicolaides K, Demers S, Hyett J, Chaillet N, Bujold E. The role of aspirin dose on the prevention of preeclampsia and fetal growth restriction: Systematic review and meta-analysis. Am J Obstet Gynecol 2017 Feb;216(2):110-120.e6. [doi: 10.1016/j.ajog.2016.09.076] [Medline: 27640943]
12. Caillon H, Tardif C, Dumontet E, Winer N, Masson D. Evaluation of sFlt-1/PlGF ratio for predicting and improving clinical management of pre-eclampsia: Experience in a specialized perinatal care center. Ann Lab Med 2018 Mar;38(2):95-101 [FREE Full text] [doi: 10.3343/alm.2018.38.2.95] [Medline: 29214752]

13. American College of Obstetricians and Gynecologists. ACOG Practice Bulletin No. 202: Gestational hypertension and preeclampsia. Obstet Gynecol 2019 Jan;133(1):e1-e25. [doi: 10.1097/AOG.0000000000003018] [Medline: 30575675]
14. Eskes M, Waelput A, Scherjon S, Bergman K, Abu-Hanna A, Ravelli A. Small for gestational age and perinatal mortality at term: An audit in a Dutch national cohort study. Eur J Obstet Gynecol Reprod Biol 2017 Aug;215:62-67. [doi: 10.1016/j.ejogrb.2017.06.002] [Medline: 28601729]
15. Ewing AC, Ellington SR, Shapiro-Mendoza CK, Barfield WD, Kourtis AP. Full-term small-for-gestational-age newborns in the US: Characteristics, trends, and morbidity. Matern Child Health J 2017 Apr;21(4):786-796 [FREE Full text] [doi: 10.1007/s10995-016-2165-z] [Medline: 27502090]
16. Wright D, Syngelaki A, Akolekar R, Poon LC, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal characteristics and medical history. Am J Obstet Gynecol 2015 Jul;213(1):62.e1-62.e10. [doi: 10.1016/j.ajog.2015.02.018] [Medline: 25724400]
17. O'Gorman N, Wright D, Syngelaki A, Akolekar R, Wright A, Poon LC, et al. Competing risks model in screening for preeclampsia by maternal factors and biomarkers at 11-13 weeks gestation. Am J Obstet Gynecol 2016 Jan;214(1):103.e1-103.e12. [doi: 10.1016/j.ajog.2015.08.034] [Medline: 26297382]
18. Gallo DM, Wright D, Casanova C, Campanero M, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal factors and biomarkers at 19-24 weeks' gestation. Am J Obstet Gynecol 2016 May;214(5):619.e1-619.e17. [doi: 10.1016/j.ajog.2015.11.016] [Medline: 26627730]
19. Tsiakkas A, Saiid Y, Wright A, Wright D, Nicolaides KH. Competing risks model in screening for preeclampsia by maternal factors and biomarkers at 30-34 weeks' gestation. Am J Obstet Gynecol 2016 Jul;215(1):87.e1-87.e17. [doi: 10.1016/j.ajog.2016.02.016] [Medline: 26875953]
20. Andrietti S, Silva M, Wright A, Wright D, Nicolaides KH. Competing-risks model in screening for pre-eclampsia by maternal factors and biomarkers at 35-37 weeks' gestation. Ultrasound Obstet Gynecol 2016 Jul;48(1):72-79 [FREE Full text] [doi: 10.1002/uog.15812] [Medline: 26566592]
21. O'Gorman N, Wright D, Poon LC, Rolnik DL, Syngelaki A, Wright A, et al. Accuracy of competing-risks model in screening for pre-eclampsia by maternal factors and biomarkers at 11-13 weeks' gestation. Ultrasound Obstet Gynecol 2017 Jun;49(6):751-755 [FREE Full text] [doi: 10.1002/uog.17399] [Medline: 28067011]
22. Nuriyeva G, Kose S, Tuna G, Kant M, Akis M, Altunyurt S, et al. A prospective study on first trimester prediction of ischemic placental diseases. Prenat Diagn 2017 Apr;37(4):341-349. [doi: 10.1002/pd.5017] [Medline: 28165141]
23. Perales A, Delgado JL, de la Calle M, García-Hernández JA, Escudero AI, Campillos JM, STEPS investigators. sFlt-1/PlGF for prediction of early-onset pre-eclampsia: STEPS (Study of Early Pre-eclampsia in Spain). Ultrasound Obstet Gynecol 2017 Sep;50(3):373-382 [FREE Full text] [doi: 10.1002/uog.17373] [Medline: 27883242]
24. Sonek J, Krantz D, Carmichael J, Downing C, Jessup K, Haidar Z, et al. First-trimester screening for early and late preeclampsia using maternal characteristics, biomarkers, and estimated placental volume. Am J Obstet Gynecol 2018 Jan;218(1):126.e1-126.e13. [doi: 10.1016/j.ajog.2017.10.024] [Medline: 29097177]
25. Tan MY, Wright D, Syngelaki A, Akolekar R, Cicero S, Janga D, et al. Comparison of diagnostic accuracy of early screening for pre-eclampsia by NICE guidelines and a method combining maternal factors and biomarkers: Results of SPREE. Ultrasound Obstet Gynecol 2018 Jun;51(6):743-750 [FREE Full text] [doi: 10.1002/uog.19039] [Medline: 29536574]
26. Wright A, Wright D, Syngelaki A, Georgantis A, Nicolaides KH. Two-stage screening for preterm preeclampsia at 11-13 weeks' gestation. Am J Obstet Gynecol 2019 Feb;220(2):197.e1-197.e11. [doi: 10.1016/j.ajog.2018.10.092] [Medline: 30414394]
27. Wright D, Tan MY, O'Gorman N, Poon LC, Syngelaki A, Wright A, et al. Predictive performance of the competing risk model in screening for preeclampsia. Am J Obstet Gynecol 2019 Feb;220(2):199.e1-199.e13. [doi: 10.1016/j.ajog.2018.11.1087] [Medline: 30447210]
28. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc 2016 Nov;23(6):1166-1173 [FREE Full text] [doi: 10.1093/jamia/ocw028] [Medline: 27174893]
29. Chicco D. Ten quick tips for machine learning in computational biology. BioData Min 2017;10:35 [FREE Full text] [doi: 10.1186/s13040-017-0155-3] [Medline: 29234465]
30. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. J Med Internet Res 2016 Dec 16;18(12):e323 [FREE Full text] [doi: 10.2196/jmir.5870] [Medline: 27986644]
31. Premru-Srsen T. Mendeley Data, v1. 2018 Jan 10. Uterine arteries Doppler and sFlt-1/PlGF ratio in hypertensive disorders during pregnancy [public dataset] URL: https://data.mendeley.com/datasets/zsjhvy9ytx/1 [accessed 2019-03-11]
32. Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. 4th edition. Cambridge, MA: Morgan Kaufmann; 2017.
33. Kotthoff L, Thornton C, Hoos HH, Hutter F, Leyton-Brown K. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. J Mach Learn Res 2017;18:1-5 [FREE Full text] [doi: 10.1145/2487575.2487629]
34. Jung Y, Hu J. A K-fold averaging cross-validation procedure. J Nonparametr Stat 2015;27(2):167-179 [FREE Full text] [doi: 10.1080/10485252.2015.1010532] [Medline: 27630515]

35. Ounpraseuth S, Lensing SY, Spencer HJ, Kodell RL. Estimating misclassification error: A closer look at cross-validation based methods. BMC Res Notes 2012 Nov 28;5:656 [FREE Full text] [doi: 10.1186/1756-0500-5-656] [Medline: 23190936]
36. Wagenmakers E, Farrell S. AIC model selection using Akaike weights. Psychon Bull Rev 2004 Feb;11(1):192-196. [doi: 10.3758/bf03206482] [Medline: 15117008]
37. Brewer M, Butler A, Cooksley S. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. Methods Ecol Evol 2016 Jun 13;7(6):679-692 [FREE Full text] [doi: 10.1111/2041-210X.12541]
38. Zeisler H, Llurba E, Chantraine F, Vatish M, Staff AC, Sennström M, et al. Predictive value of the sFlt-1:PlGF ratio in women with suspected preeclampsia. N Engl J Med 2016 Jan 07;374(1):13-22. [doi: 10.1056/NEJMoa1414838] [Medline: 26735990]
39. Sabrià E, Lequerica-Fernández P, Ganuza PL, Ángeles EE, Escudero AI, Martínez-Morillo E, et al. Use of the sFlt-1/PlGF ratio to rule out preeclampsia requiring delivery in women with suspected disease. Is the evidence reproducible? Clin Chem Lab Med 2018 Jan 26;56(2):303-311. [doi: 10.1515/cclm-2017-0443] [Medline: 28841572]
40. Chen Q, Izumi A, Minakami H, Sato I. Comparative changes in uterine artery blood flow waveforms in singleton and twin pregnancies. Gynecol Obstet Invest 1998;45(3):165-169. [doi: 10.1159/000009948] [Medline: 9565139]
41. Frusca T, Gervasi M, Paolini D, Dionisi M, Ferre F, Cetin I. Budget impact analysis of sFlt-1/PlGF ratio as prediction test in Italian women with suspected preeclampsia. J Matern Fetal Neonatal Med 2017 Sep;30(18):2166-2173. [doi: 10.1080/14767058.2016.1242122] [Medline: 27737599]
42. Shao Y, Qiu J, Huang H, Mao B, Dai W, He X, et al. Pre-pregnancy BMI, gestational weight gain and risk of preeclampsia: A birth cohort study in Lanzhou, China. BMC Pregnancy Childbirth 2017 Dec 01;17(1):400 [FREE Full text] [doi: 10.1186/s12884-017-1567-2] [Medline: 29191156]
43. Afrakhteh M, Moeini A, Taheri MS, Haghighatkhah HR, Fakhri M, Masoom N. Uterine Doppler velocimetry of the uterine arteries in the second and third trimesters for the prediction of gestational outcome. Rev Bras Ginecol Obstet 2014 Jan;36(1):35-39. [doi: 10.1590/S0100-72032014000100008] [Medline: 24554228]
44. Tarasevičienė V, Grybauskienė R, Mačiulevičienė R. sFlt-1, PlGF, sFlt-1/PlGF ratio and uterine artery Doppler for preeclampsia diagnostics. Medicina (Kaunas) 2016;52(6):349-353 [FREE Full text] [doi: 10.1016/j.medici.2016.11.008] [Medline: 27940029]
45. Rizos D, Eleftheriades M, Karampas G, Rizou M, Haliassos A, Hassiakos D, et al. Placental growth factor and soluble fms-like tyrosine kinase-1 are useful markers for the prediction of preeclampsia but not for small for gestational age neonates: A longitudinal study. Eur J Obstet Gynecol Reprod Biol 2013 Dec;171(2):225-230. [doi: 10.1016/j.ejogrb.2013.08.040] [Medline: 24035323]
46. Albu AR, Anca AF, Horhoianu VV, Horhoianu IA. Predictive factors for intrauterine growth restriction. J Med Life 2014 Jun 15;7(2):165-171 [FREE Full text] [Medline: 25408721]
47. Kwiatkowski S, Bednarek-Jędrzejek M, Ksel J, Tousty P, Kwiatkowska E, Cymbaluk A, et al. sFlt-1/PlGF and Doppler ultrasound parameters in SGA pregnancies with confirmed neonatal birth weight below 10th percentile. Pregnancy Hypertens 2018 Oct;14:79-85. [doi: 10.1016/j.preghy.2018.08.448] [Medline: 30527123]
48. Contro E, Maroni E, Cera E, Youssef A, Bellussi F, Pilu G, et al. Unilaterally increased uterine artery resistance, placental location and pregnancy outcome. Eur J Obstet Gynecol Reprod Biol 2010 Dec;153(2):143-147. [doi: 10.1016/j.ejogrb.2010.07.012] [Medline: 20667646]
49. Michael G, Kumaravel A, Chandrasekar A. Detection of malicious attacks by meta classification algorithms. Int J Adv Netw Appl 2015;6(5):2455-2459 [FREE Full text]
50. Podgorelec V, Kokol P, Stiglic B, Rozman I. Decision trees: An overview and their use in medicine. J Med Syst 2002 Oct;26(5):445-463. [doi: 10.1023/a:1016409317640] [Medline: 12182209]
51. Lin L, Wang Q, Sadek AW. A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations. Accid Anal Prev 2016 Jun;91:114-126. [doi: 10.1016/j.aap.2016.03.001] [Medline: 26974028]
52. Ergin RN, Yayla M. Uterine artery pulsatility index and diastolic notch laterality according to the placental location. Clin Exp Obstet Gynecol 2015;42(5):640-643. [Medline: 26524814]
53. Poon LC, Staboulidou I, Maiz N, Plasencia W, Nicolaides KH. Hypertensive disorders in pregnancy: Screening by uterine artery Doppler at 11-13 weeks. Ultrasound Obstet Gynecol 2009 Aug;34(2):142-148 [FREE Full text] [doi: 10.1002/uog.6452] [Medline: 19644947]
54. Hoogland HJ, de Haan J. Ultrasonographic placental localization with respect to fetal position in utero. Eur J Obstet Gynecol Reprod Biol 1980 Sep;11(1):9-15. [doi: 10.1016/0028-2243(80)90047-7] [Medline: 7193612]
55. Koken GN, Kanat-Pektas M, Kayman Köse S, Arioz DT, Yilmazer M. Maternal blood pressure and dominant sleeping position may affect placental localization. J Matern Fetal Neonatal Med 2014 Oct;27(15):1564-1567. [doi: 10.3109/14767058.2013.870547] [Medline: 24283300]
56. Bartsch E, Medcalf KE, Park AL, Ray JG, High Risk of Pre-eclampsia Identification Group. Clinical risk factors for pre-eclampsia determined in early pregnancy: Systematic review and meta-analysis of large cohort studies. BMJ 2016 Apr 19;353:i1753 [FREE Full text] [doi: 10.1136/bmj.i1753] [Medline: 27094586]
57. Zhang J, Simonti CN, Capra JA. Genome-wide maps of distal gene regulatory enhancers active in the human placenta. PLoS One 2018;13(12):e0209611 [FREE Full text] [doi: 10.1371/journal.pone.0209611] [Medline: 30589856]

Abbreviations

AIC: Akaike information criterion
AIC_C: corrected Akaike information criterion
AUC: area under the receiver operating characteristic curve
CVR: classification via regression
FN: false negative
FP: false positive
IR: information retrieval
IUGR: intrauterine growth restriction
LM: linear model
MAP: mean arterial pressure
MCC: Matthew correlation coefficient
MOE: Ministry of Education
MOST: Ministry of Science and Technology
N: negative
P: positive
PDD: placental dysfunction–related disorder
PI: pulsatility index
PI-UtA: pulsatility index of the uterine artery
PlGF: placental growth factor
PPV: positive predictive value
PRC: precision-recall curve
PSV: peak systolic velocity
PSV-UtA: peak systolic velocity of the uterine artery
RI: resistivity index
RI-UtA: resistivity index of the uterine artery
RMSE: root mean square error
ROC: receiver operating characteristic
sFlt-1: soluble fms-like tyrosine kinase receptor-1
STROBE: STrengthening the Reporting of OBservational studies in Epidemiology
TN: true negative
TP: true positive
UtA: uterine artery
Weka: Waikato Environment for Knowledge Analysis

Edited by G Eysenbach; submitted 10.07.19; peer-reviewed by J Rezende Filho, G Borgulya; comments to author 21.10.19; revised version received 11.11.19; accepted 23.03.20; published 18.05.20.

Please cite as:
Sufriyana H, Wu YW, Su ECY
Prediction of Preeclampsia and Intrauterine Growth Restriction: Development of Machine Learning Models on a Prospective Cohort
JMIR Med Inform 2020;8(5):e15411
URL: http://medinform.jmir.org/2020/5/e15411/
doi:10.2196/15411
PMID:32348266

Original Paper

Multi-Level Representation Learning for Chinese Medical Entity Recognition: Model Development and Validation

Zhichang Zhang[1], PhD; Lin Zhu[1], MS; Peilin Yu[1], MS

College of Computer Science and Engineering, University of Northwest Normal, Lanzhou, China

Corresponding Author:
Zhichang Zhang, PhD
College of Computer Science and Engineering
University of Northwest Normal
967 Anning East Road
Lanzhou,
China
Phone: 86 13038769329
Email: zzc@nwnu.edu.cn

Abstract

Background: Medical entity recognition is a key technology that supports the development of smart medicine. Existing methods on English medical entity recognition have undergone great development, but their progress in the Chinese language has been slow. Because of limitations due to the complexity of the Chinese language and annotated corpora, these methods are based on simple neural networks, which cannot effectively extract the deep semantic representations of electronic medical records (EMRs) and be used on the scarce medical corpora. We thus developed a new Chinese EMR (CEMR) dataset with six types of entities and proposed a multi-level representation learning model based on Bidirectional Encoder Representation from Transformers (BERT) for Chinese medical entity recognition.

Objective: This study aimed to improve the performance of the language model by having it learn multi-level representation and recognize Chinese medical entities.

Methods: In this paper, the pretraining language representation model was investigated; utilizing information not only from the final layer but from intermediate layers was found to affect the performance of the Chinese medical entity recognition task. Therefore, we proposed a multi-level representation learning model for entity recognition in Chinese EMRs. Specifically, we first used the BERT language model to extract semantic representations. Then, the multi-head attention mechanism was leveraged to automatically extract deeper semantic information from each layer. Finally, semantic representations from multi-level representation extraction were utilized as the final semantic context embedding for each token and we used softmax to predict the entity tags.

Results: The best F1 score reached by the experiment was 82.11% when using the CEMR dataset, and the F1 score when using the CCKS (China Conference on Knowledge Graph and Semantic Computing) 2018 benchmark dataset further increased to 83.18%. Various comparative experiments showed that our proposed method outperforms methods from previous work and performs as a new state-of-the-art method.

Conclusions: The multi-level representation learning model is proposed as a method to perform the Chinese EMRs entity recognition task. Experiments on two clinical datasets demonstrate the usefulness of using the multi-head attention mechanism to extract multi-level representation as part of the language model.

(JMIR Med Inform 2020;8(5):e17637) doi:10.2196/17637

KEYWORDS

medical entity recognition; multi-level representation learning; Chinese; natural language processing; electronic medical records; multi-head attention mechanism

Introduction

Background

Electronic medical records (EMRs) comprise patients' health information. Diagnostic accuracy can be improved by making full use of the available information in EMRs. Medical entity recognition (ER) is a fundamental task of medical natural language processing (NLP) and is usually treated as a sequence labeling problem [1]. As shown in Figure 1, in which three predefined entity categories are disease, drug, and treatment, when using the BIO (beginning of the noun phrase, middle of the noun phrase, and not a noun phrase) labeling mode to tag Chinese EMRs, the candidate label set contains seven types: B-Dis (disease), I-Dis, B-Med (medicine), I-Med, B-Tre (treatment), I-Tre, and O.

Figure 1. A tagging example of Chinese electronic medical records.

Sentence	Tagging Results
临床初步诊断：急性支气管炎，口服磺啶新林胶囊。给予抗炎、解痉、祛痰对症支持治疗，完善相关检查。 (The preliminary analysis of the clinical diagnosis : acute bronchitis, oral sulfamethoxazole capsules . Give anti-inflammatory, antispasmodic, expectorant symptomatic support treatment, perfect relevant examination.)	临/O床/O初/O步/O诊/O断/O：/O急/B_Dis性/I_Dis支/I_Dis气/I_Dis管/I_Dis炎/I_Dis，/O口/O服/O磺/B_Med啶/I_Med新/I_Med林/I_Med胶/I_Med囊/I_Med。/O给/O予/O抗/B_Tre炎/I_Tre、/O解/B_Tre痉/I_Tre、/O祛/B_Tre痰/I_Tre对/B_Tre症/I_Tre支/I_Tre持/I_Tre治/O疗/O，/O完/O善/O相/O关/O检/O查/O。/O

Generally, the methods of ER can be divided into two categories. The first category leverages rules and dictionaries to represent linguistic features and domain knowledge to identify clinical entities [2]. The second category is based on traditional machine learning and neural networks [3-8]; this type of method greatly improves the performance of ER models but requires large-scale labeled data during model parameter training. In the medical field, creating annotation datasets is restricted by professional knowledge and legal regulations, so the lack of annotated corpora becomes one of the greatest technical challenges. At present, ER attracts a lot of attention from the field to improve the representation learning capability of current methods. Research studies have demonstrated that using embedding techniques can help solve the problem of missing supervised data in NLP tasks, including the factorization methods of Global Vectors (GloVe) [9], the neural methods of word2vec [10] and fastText [11], and more recent dynamic methods that take into account the context, such as Embeddings from Language Models (ELMo) [12] and OpenAI Generative Pre-trained Transformer (GPT) [13]. Those embedding technologies can capture the context of semantics in unsupervised data and generate different vector representations of the same word in different contextual situations.

Among them, Bidirectional Encoder Representations from Transformers (BERT) [14] integrates many top ideas of language models and gives a particularly prominent performance. Transform-block is a feature extractor and learns different types of abstract granularity information. Multi-layer information is iterated layer by layer to generate embedding representation. In the actual training process, most downstream tasks take BERT's last embedding vector as the input of the model. However, studies found that different NLP tasks have different characteristics of requirements. Therefore, combining task features into the language model can reduce the loss of extracted information by the feature extractor and improve the utilization of language models. For example, Peters et al [12] explicitly showed that the lower layer fits into the local semantic relationships, the higher layer is suitable for longer-range relationships, and the final layer specializes in the language model. Peters et al [15] also showed that combining all semantic internal states models, by using a weighted-sum method to represent the vector of a word, can enrich the characteristics of the word in learning deep contextualized embedding representations. Because the Chinese ER task focuses on word granularity information, this is a straightforward way to use the information extracted from the low-layer representation.

In this work, we tackle representation using the BERT language model. Our objective is to extract each layer of semantic information using feature extractors. We constructed a multi-level representation learning model for the optimal integration of information. Our contributions can be summarized as follows:

1. We manually annotated a new Chinese EMR (CEMR) corpus for ER tasks. Moreover, we propose a multi-level representation learning model to mine hidden representation.
2. The proposed model takes advantage of the multi-head attention mechanism to integrate more suitable information from each layer and can perform as a state-of-the-art method on two clinical text datasets.
3. The best F1 score achieved by the experiment was 82.11% on the CEMR corpus and significant improvement on the CCKS (China Conference on Knowledge Graph and Semantic Computing) 2018 benchmark dataset was attained.

Chinese Electronic Medical Record Dataset: A Newly Constructed Corpus

Large labeled datasets are not always readily accessible. To facilitate the research on the ER task of the Chinese EMRs and

future work in related topics, we constructed a new manually annotated CEMR dataset. The normalization of the labeling process refers to a large number of annotation guidelines [16]. All EMRs came from Third-Class A-Level hospitals in Gansu Province, China, which contained 80,000 EMRs across 14 departments. Manual labeling of 4000 medical records provided the data for ER experiments. Table 1 shows the data distribution of the 14 hospital departments. The CEMR corpus contains six types of entities: disease (Dis), symptom (Sym), test, treatment (Tre), medicine (Med), and abnormal inspection result (Abn). The categories are defined as follows:

1. Disease: refers to a specific abnormal pathological condition. This abnormal life condition is caused by disorders of self-regulation, such as diabetes.
2. Symptom: refers to subjective feelings described by patients or objective facts observed externally, such as abdominal distension.
3. Test: includes examination procedures, items, and equipment to collect and confirm more information about the disease or symptom, such as electrocardiogram.
4. Treatment: refers to a treatment program or intervention to treat diseases or relieve symptoms, such as neurotrophic treatment.
5. Medicine: refers to a chemical substance used to prevent and treat diseases or to strengthen the body and improve mental state, such as insulin.
6. Abnormal inspection result: refers to an abnormal change or inspection result observed by doctors or by examination equipment, such as a little sputum sound.

Before labeling the data, private information was removed in the EMRs, such as patients' names, addresses, and hospital IDs. In the process of labeling samples, the annotation tool is developed specifically for the ER task. Moreover, some strategies have been developed to create high-quality annotated data. For example, the annotation samples will be randomly checked at any time.

Table 1. Electronic medical record (EMR) data distribution by department.

Department	EMR count, n (%)
Neurosurgery	77 (1.93)
Neurology	77 (1.93)
Cardiology	77 (1.93)
Gynecology and obstetrics	77 (1.93)
Andrology	77 (1.93)
Respiratory medicine	77 (1.93)
Cardiovasology	77 (1.93)
Hepatobiliary surgery	77 (1.93)
Ophthalmology	77 (1.93)
Orthopedics	77 (1.93)
Gynecology	101 (2.53)
Pediatrics	232 (5.80)
Internal medicine	970 (24.25)
Surgery	1495 (37.38)
Other	432 (10.80)
Total	4000 (100)

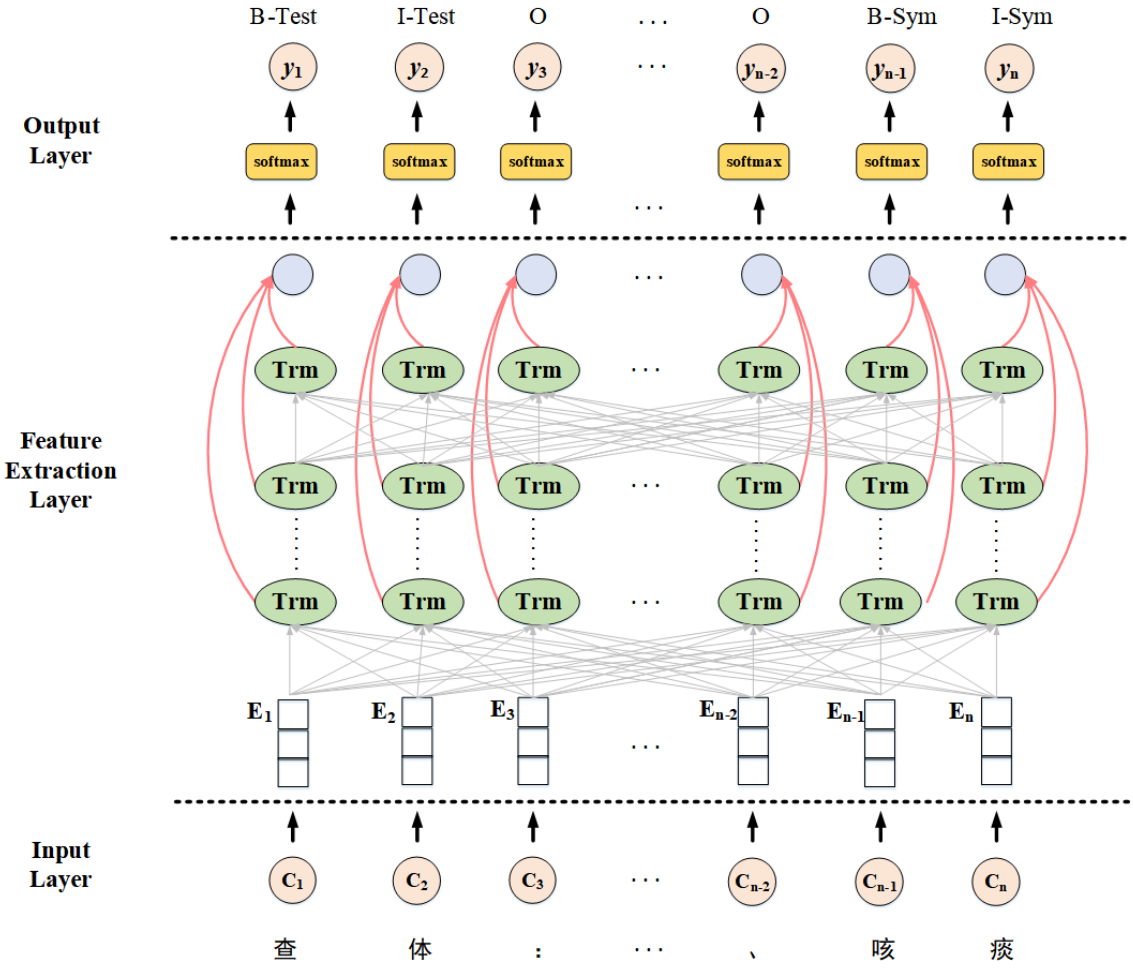
Methods

Overview

The goal of the ER task is to provide the model with an EMR and its semantic types, so that it can extract and classify all characters in the text. The proposed model consists of three stacked layers: the input layer, the feature extraction layer, and the output layer.

As shown in Figure 2, the model first used the BERT language model to extract the semantic representations. Then, the multi-head attention mechanism was leveraged to automatically extract deeper semantic information from each layer. Finally, the semantic information from the multi-level representation extraction was utilized as the final semantic context embedding for each token and was input into the softmax layer to predict the entity tag. The input sentence was denoted as $C = (c_1, c_2, c_3, ..., c_n)$, where c_n represented the *n*-th character in sentence C of the Chinese EMR. Correspondingly, the output sentence's predicted tag sequence was denoted as $Y = (y_1, y_2, y_3 ..., y_n)$, where y_n belonged to one of the sets: B-Dis, I-Dis, B-Sym, I-Sym, B-Test, I-Test, B-Tre, I-Tre, B-Med, I-Med, B-Abn, I-Abn, or O. In the following text, we introduce the BERT language model and describe the proposed multi-level representation learning model.

Figure 2. Multi-level representation learning for ER model. B-Sym: beginning of the noun phrase for the symptom entity; B-Test: beginning of the noun phrase for the test entity; C: input sentence; E: input embedding; I-Sym: middle of the noun phrase for the symptom entity; I-Test: middle of the noun phrase for the test entity; O: not a noun phrase; Trm: transform-block; y: output sentence's predicted tag sequence.



Bidirectional Encoder Representations From Transformers

BERT was designed to learn deep bidirectional representations by jointly conditioning both the left and right contexts in all layers. It was based on multi-layer bidirectional encoder transformers and could be used for different architectures. When given a character-level sequence C = (c_1, c_2, c_3, ..., c_n), BERT was formulated as follows:

$$h_1 = E_{Token} + E_{Segment} + E_{Position} \quad (1)$$

$$h_l = Trm(h_{l-1}) \quad (2)$$

$$Y^{BERT} = Softmax(w_O h_L + b_O) \quad (3)$$

where h_1 represents input embedding for a sequence and is made up of E_{Token}, $E_{Segment}$, and $E_{Position}$, which mean token, segment, and position for a sentence, respectively. The BERT leverage transformer is the feature extractor. *Trm* is a transform-block that includes self-attention, the fully connected layers, and the output layer. The current *l* layer hidden state came from the upper *l–1* layer and *L* was the last layer. Y^{BERT} denotes the output layer that predicts the sequence labels. In the above equations, w_O denotes the function weight and b_O is the function bias. All parameters of the transform-block were trained in advance on a large document-level corpus using a masked language model and were fine-tuned by predicting task-specific labels with the output layer to maximize the log-probability of the correct label.

Multi-Level Representation Learning for Entity Recognition

The Multi-Level Representation Learning for ER model (Multi-Level ER) could automatically integrate deeper semantic information from all layers of the feature extractor for ER task. The proposed language model took advantage of the multi-head attention mechanism. Multi-head attention is a special type of attention that allowed the model to focus on different positions of subspace representation information and could learn more about the connections between internal elements. Figure 3 shows the calculation process of the multi-head attention mechanism when calculating the weight of the transform-block output knowledge. The query (Q), key (K), and value (V) in the transform-block were calculated. The process of acquiring Q, K, and V could be written as follows:

$$H = Concat(h_1, h_2, h_3, ..., h_L) \quad (4)$$

$$Q = w_Q h_L + b_L \quad (5)$$

$$K = w_K H + b_K \quad (6)$$

$$V = w_V H + b_V \quad (7)$$

where h_L denotes the hidden state of the final layer of the transform-block. The parameters w_Q, w_K, and w_V are weight matrices. The parameters b_L, b_K, and b_V are bias matrices. The attention function is calculated as follows:

$$head_i = Softmax(Q_L K^T/\sqrt{d})V \quad (8)$$

where $head_i$ means the i-th head. Q_L is the query key value of the last L layer. $\sqrt{d}$ is used to control the order of magnitude of calculation results and d donates the dimension of the K vector.

In this work, we used multi-head attention, as introduced in the following equation:

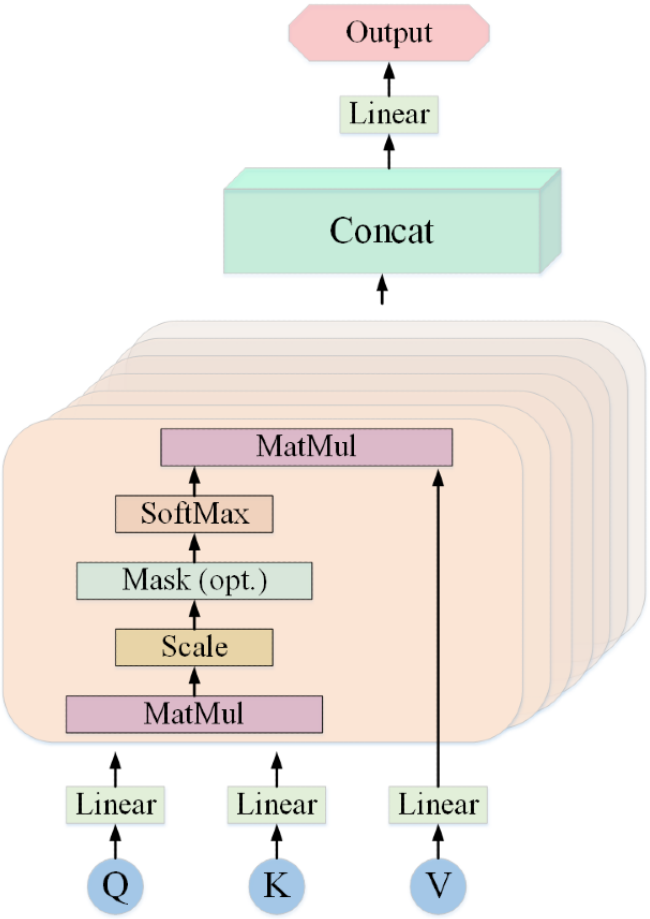
$$E = Concat(head_1, head_2, head_3, ..., head_l)w_h + b_h \quad (9)$$

where w_h is used to balance the head weight. For the final layer of the network, we pass the results into a fully connected layer with a softmax function, as follows:

$$Y^{Multi\text{-}Level\ ER} = Softmax(w_O E + b_O) \quad (10)$$

where w_O is the output weight matrix and b_O is the bias of the output layer.

Figure 3. Multi-head attention mechanism. K: key; Q: query; V: value.



Experiment

This model was supported by multiple sets of comparative experiments. Each group of experiments was repeated three times, and the result in the middle of the ranking was taken as the final result.

Dataset and Evaluation Criteria

We evaluated the proposed model on two datasets: the CCKS 2018 dataset and the CEMR dataset. The CCKS 2018 dataset was adopted from the Chinese EMR named ER task at the CCKS, which included 1000 admission records. In the experiment, 600 records were used as training data and the remaining were test data. Comparative experiments were made on the new CEMR corpus and contained 4000 documents. We further split the corpus set by 60%, 20%, and 20% as training, validation, and test sets, respectively. Table 2 shows the distribution of documents in two datasets.

Table 2. Components of the two datasets.

Dataset	Number of records per set			
	Total	Training set	Validation set	Test set
CEMR[a] dataset	4000	2400	800	800
CCKS[b] 2018	1000	600	N/A[c]	400

[a]CEMR: Chinese electronic medical record.

[b]CCKS: China Conference on Knowledge Graph and Semantic Computing.

[c]Not applicable; because the comparison method does not divide the validation set on the CCKS dataset, we have kept this the same as the original experiment to make the comparison fair.

To evaluate the performance of all prediction methods fairly, the results were validated by precision (P), recall (R), and F1 scores (F1) as measurements to evaluate the recognition effectiveness of the model; these were defined as follows:

$$P = TP/(TP + TF) \quad (11)$$

$$R = TP/(TP + FN) \quad (12)$$

$$F1 = (2 \times T \times P)/(P + R) \quad (13)$$

An entity is annotated as correct when its category and boundary are fully labeled correctly. TP is the count of entity labels presenting the same labels as gold standard labels, FP is the count of recognized entities marked incorrectly in the results, and FN is the count of the gold standard entities that are not present in the results of the indicator.

Parameter Setup

Hyperparameter configuration was adjusted according to the performance on the described validation sets. We used a publicly available pretraining language representation model, namely the $BERT_{BASE\text{-}Chinese\text{-}uncased}$. This model has 12 layers, 768 hidden layers, and 12 heads. The multi-head attention mechanism was utilized to automatically integrate all layers of information. By comparing experimental results with different head numbers, we had set the head number to 12. We fine-tuned the model over 10 epochs with a batch size of 32. The maximum training sentence length was 64. The model was trained with the AdamW optimizer with a learning rate of le-5 and we applied a dropout rate of 0.3.

Results

Overview

We summarized the overall performance by computing the F1 score; the results are illustrated in Table 3. On the CEMR dataset, we compared the multi-level ER learning model with previous classic methods, including conditional random field (CRF), convolutional neural network (CNN)+bidirectional long short-term memory (BiLSTM)+CRF, lattice long short-term memory (LSTM), and BERT. We found that the proposed model is better than state-of-the-art baseline methods, with F1 scores of 0.94% to 4.9%. Our multi-level ER learning model had improved by 1.48% in its *P* value, 0.47% in its R value, and 0.94% in its F1 score compared to the BERT model. The result also demonstrated that pretraining the multi-level ER learning language model was highly effective for task-specific Chinese EMR ER.

Table 3. Comparison of method performance on the Chinese electronic medical record (CEMR) dataset.

Method	*P* value (%)	R value (%)	F1 score (%)
Conditional random field (CRF)	88.57	68.43	77.21
CNN[a]+BiLSTM[b]+CRF	81.51	76.92	79.15
Lattice long short-term memory (LSTM)	88.60	74.48	80.93
Bidirectional Encoder Representations from Transformers (BERT)	83.73	78.76	81.17
Multi-level representation learning for entity recognition (multi-level ER)	85.21	79.23	82.11

[a]CNN: convolutional neural network.

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

We also applied our model to the widely used benchmark CCKS 2018 dataset and used the same data split to compare it. Huang et al [17] proposed a BiLSTM-CRF model for sequence tagging and Cai et al [18] was based on the self-matching attention mechanism (SM) and proposed an SM-LSTM-CRF model design for the named ER task. The results are shown in Table 4. Under the condition of not needing any external resources, the proposed multi-level ER learning model already outperformed the previous SM-LSTM-CRF model by 3.1% on the F1 score.

Table 4. Comparison of method performance on the China Conference on Knowledge Graph and Semantic Computing 2018 dataset.

Method	*P* value (%)	R value (%)	F1 score (%)
BiLSTM[a]-CRF[b] [17]	65.68	69.04	67.32
SM[c]-LSTM-CRF [18]	80.54	79.61	80.08
Multi-level representation learning for entity recognition (multi-level ER)	83.90	82.47	83.18

[a]BiLSTM: bidirectional long short-term memory.

[b]CRF: conditional random field.

[c]SM: self-matching attention mechanism.

The Effect of Assembling Methods

We compared the effects of different assembling methods on model performance to verify the ability of the multi-head attention mechanism to combine hierarchical information. As listed in Table 5, we first applied concatenation that directed the horizontal concatenated tensors; the F1 score was 81.51%. We then adopted the sum average method to get an F1 score of 81.11%. We finally adopted the multi-head attention method, given that it had the best overall performance compared to several other methods we evaluated. The results showed that integrated hidden information can acquire more suitable representation; the multi-head attention mechanism can be leveraged to automatically extract deeper semantic information from each layer, which is the most effective assembling method.

Table 5. The effect of assembling methods.

Assembling method	*P* value (%)	R value (%)	F1 score (%)
Concatenation	84.22	78.97	81.51
Sum average	83.27	79.06	81.11
Multi-head attention mechanism	85.21	79.23	82.11

The Effect of Extraction Layer Numbers

To examine the impact of extraction layer numbers on model performance, we performed comparative experiments using various extraction layer numbers; the results are shown in Table 6. It was observed that the performance of all layers was superior to that of the other numbers of layers, which introduced multi-level ER into the language model and enhanced model performance. By and large, the tendency was that performance improved as the number of extracting layers increased. However, we also discovered that extracting the last four layers gave higher F1 scores than extracting the last six or two layers. The analysis showed that the results were closely related to the specific dataset. Of course, as the number of layers increased, parameters required by the neural network also increased significantly. Therefore, when there was a high demand for speed on the model, we could select a structure that included the last four layers to optimize time efficiency.

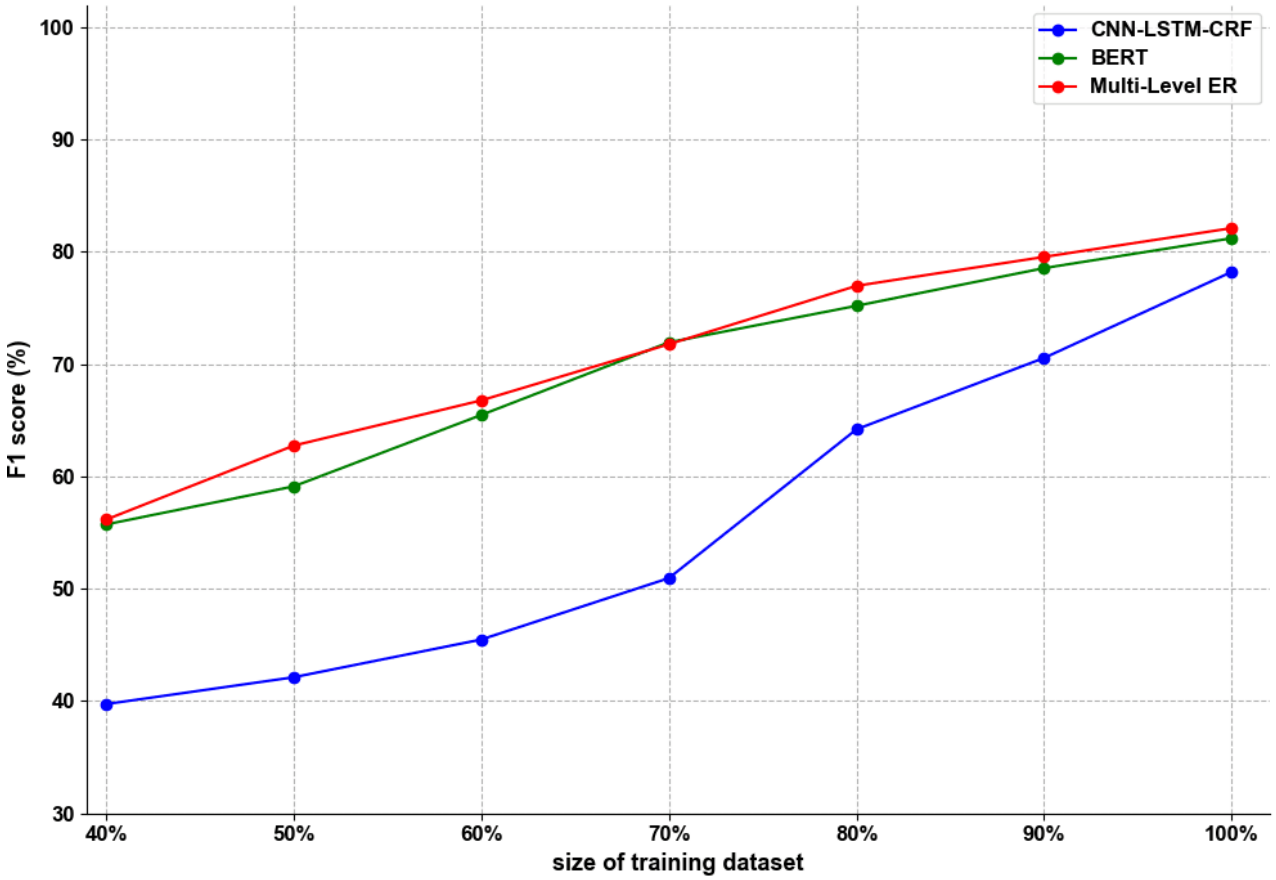
Table 6. The effect of extracted layer numbers.

Extraction layer number	*P* value (%)	R value (%)	F1 score (%)
Total layers	85.21	79.23	82.11
The last six layers	85.15	78.65	81.77
The last four layers	85.50	78.68	81.95
The last two layers	84.51	78.68	81.49

The Effect of Dataset Size

Figure 4 shows the impact of the dataset size on model performance. Horizontal coordinates represent the size of the training dataset and vertical coordinates indicate the F1 scores. During the experiment, we used different sized corpora to train the CNN-LSTM-CRF, BERT, and multi-level ER models. The figure shows that as the training dataset increased, the performance of the models also improved. In reality, we had a limited number of datasets, and models were unlikely to reach saturation. Therefore, the impact of dataset size on performance was particularly critical. We found that the CNN-LSTM-CRF model performance was sharply affected by the size of the dataset when the training set increased from 70% to 100%. Inversely, the BERT model and the multi-level ER model were less influenced by the training dataset size, and our proposed multi-level RE model outperformed the BERT model.

Figure 4. The effect of dataset size. BERT: Bidirectional Encoder Representations from Transformers; CNN: convolutional neural network; CRF: conditional random field; LSTM: long short-term memory; Multi-Level ER: multi-level representation learning for entity recognition.



Discussion

Case Studies

To show that our model was able to solve the challenge of integrating representation information, three case studies comparing the multi-level ER model with the BERT model are shown in Figure 5. Several obvious trends emerged from the comparative experiments. Most generally, when the word "disease" is included within the medical history, it is mistaken for a disease. For example, case study 1 in Figure 5 shows that "history of mental disease" is recognized as a disease. Case study 2 in Figure 5 shows that when "anal" and "external genitals" appear together before the examination, the system will only identify the adjacent area to be tested. The descriptions with the obvious word "treatment" are identified as a treatment in case study 3 of Figure 5.

Figure 5. Case studies comparing the multi-level representation learning for entity recognition (Multi-Level ER) model with the Bidirectional Encoder Representations from Transformers (BERT) model.

Case	BERT	Multi-Level ER	Correct Entity
Case 1	否认糖尿病、脑血管疾病、精神疾病史。	否认糖尿病、脑血管疾病、精神疾病史。	否认糖尿病、脑血管疾病、精神疾病史。
	Deny diabetes, cerebrovascular disease, and history of mental disease.	Deny diabetes, cerebrovascular disease, and history of mental disease.	Deny diabetes, cerebrovascular disease, and history of mental disease.
Case 2	肛门及外生殖器检查结果：无异常。	肛门及外生殖器检查结果：无异常。	肛门及外生殖器检查结果：无异常。
	Anal and external genital examination results: no abnormal.	Anal and external genital examination results: no abnormal.	Anal and external genital examination results: no abnormal.
Case 3	今患者及家属为求进一步治疗。	今患者及家属为求进一步治疗。	今患者及家属为求进一步治疗。
	The patient and his family sought further treatment.	The patient and his family sought further treatment.	The patient and his family sought further treatment.

We found that the BERT model's embedding technology improves the performance of the ER model in Chinese EMRs; however, using information from only the last layer of the feature extractor in the language model did not achieve the best experimental results. Our proposed multi-level ER model combines the information from each layer of the feature extractor and selects the most suitable, long-term, syntactic, relationship information for the ER task, which greatly improves the performance of the model.

Related Work

ER tasks attract a large amount of scholastic attention. The development of deep learning methods has resulted in a breakthrough regarding these tasks. CNN and recurrent neural network (RNN) models have emerged one after another; the attention mechanism and transfer learning were applied to the model. Wu et al [19] utilized a CNN model to generate features represented by several global hidden nodes. Both local features and global features were then fed into a standard affine network to recognize named entities in clinical text. Ju et al [20] used an LSTM neural model to identify nested entities by dynamically stacking flat, named ER layers. Rei et al [21] applied the attention mechanism to dynamically decide how much information to use from a character-level or word-level component in an end-to-end model. Lee et al [22] applied transfer learning in named ER by training a model on source task and using the trained model on the target task for fine-tuning. Peng et al [23] proposed a method where the prediction model was based on BiLSTM, which was taken as the source task of transfer learning. For the ER task in clinical notes, Bharadwaj et al's [24] work centered on effectively adapting these neural architectures toward low-resource settings using parameter transfer methods.

Language models can capture the syntactic and semantic information of words from a large number of unlabeled texts, which alleviates the problem of an insufficiently annotated corpus in special domains. Peters et al [12] used a language model to obtain a deep contextualized word pretraining representation called ELMo and improved the accuracy of six NLP tasks. Radford et al [13] proposed the GPT for language understanding tasks. For text classification and sequence labeling tasks, the transfer ability is better. Devlin et al [14] proposed the pretraining of deep bidirectional transformers for language understanding (ie, BERT); it captured true directional context information, sweeping 11 NLP tasks through pretraining and fine-tuning.

Our motivation is to seize the optimal information from each layer of a feature extractor to suit a given task. Takase et al [25] employed intermediate layer representation, including input embedding, to calculate the probability distributions to solve a ranking problem in language generation tasks. Kaneko et al [26] demonstrated that learning suitable representation came from different layers in grammatical error detection tasks. Therefore, we tracked their work and found the issue in the ER task in Chinese EMRs.

Conclusions

We propose a novel, multi-level, representation learning model for ER of Chinese EMRs-the multi-level ER model. We compared our model with state-of-the-art models and observed comparable performance without any external syntactic tools. The results showed that the use of the multi-head attention mechanism can effectively integrate deep semantic information from each layer of the feature extractor. In the future, we plan to apply multi-level ER to other language representation models in order to obtain even greater improvement.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61762081, No. 61662067, and No. 61662068) and the Key Research and Development Project of Gansu Province (No. 17YF1GA016). The datasets used and analyzed during this study are available from the first author upon reasonable request. The CCKS 2018 dataset that supports the findings of this study were adopted from the Chinese EMR named ER task from the CCKS 2018, but restrictions apply to the availability of these data, which were used under license for this study and are not publicly available.

Conflicts of Interest

None declared.

References

1. Zhao H, Yang Y, Zhang Q, Si L. Improve neural entity recognition via multi-task data selection and constrained decoding. In: Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). 2018 Presented at: 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018); June 1-6, 2018; New Orleans, LA p. 346-351 URL: https://www.aclweb.org/anthology/N18-2056.pdf [doi: 10.18653/v1/N18-2056]
2. Jiang M, Sanger T, Liu X. Combining contextualized embeddings and prior knowledge for clinical named entity recognition: Evaluation study. JMIR Med Inform 2019 Nov 13;7(4):e14850 [FREE Full text] [doi: 10.2196/14850] [Medline: 31719024]
3. Lample G, Ballesteros M, Subramanian S. Neural architectures for named entity recognition. In: Proceedings of the 2016 North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016). 2016 Presented at: 2016 North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2016); June 12-17, 2016; San Diego, CA p. 260-270 URL: https://www.aclweb.org/anthology/N16-1030.pdf [doi: 10.18653/v1/N16-1030]
4. Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016 Presented at: 54th Annual Meeting of the Association for

Computational Linguistics; August 7-12, 2016; Berlin, Germany p. 1064-1074 URL: https://www.aclweb.org/anthology/P16-1101.pdf [doi: 10.18653/v1/P16-1101]
5. Yang Z, Salakhutdinov R, William WC. Transfer learning for sequence tagging with hierarchical recurrent networks. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017). 2017 Presented at: 5th International Conference on Learning Representations (ICLR 2017); April 24-26, 2017; Toulon, France p. 1-10 URL: https://arxiv.org/pdf/1703.06345.pdf
6. Lee K, He L, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017 Presented at: 2017 Conference on Empirical Methods in Natural Language Processing; September 7-11, 2017; Copenhagen, Denmark p. 188-197 URL: https://www.aclweb.org/anthology/D17-1018.pdf [doi: 10.18653/v1/D17-1018]
7. Chen X, Shi Z, Qiu X, Huang X. Adversarial multi-criteria learning for Chinese word segmentation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017 Presented at: 55th Annual Meeting of the Association for Computational Linguistics; July 30-August 4, 2017; Vancouver, Canada p. 1193-1203 URL: https://www.aclweb.org/anthology/P17-1110.pdf [doi: 10.18653/v1/P17-1110]
8. El Boukkouri H, Ferret O, Lavergne T, Zweigenbaum P. Embedding strategies for specialized domains: Application to clinical entity recognition. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. 2019 Presented at: 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop; July 28-August 2, 2019; Florence, Italy p. 295-301 URL: https://www.aclweb.org/anthology/P19-2041.pdf [doi: 10.18653/v1/P19-2041]
9. Pennington J, Socher R, Manning C. GloVe: Global Vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014 Presented at: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); October 25-29, 2014; Doha, Qatar p. 1532-1543 URL: https://www.aclweb.org/anthology/D14-1162.pdf
10. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations 2013. 2013 Presented at: International Conference on Learning Representations 2013; May 2-4, 2013; Scottsdale, Arizona p. 1-12 URL: https://arxiv.org/pdf/1301.3781.pdf
11. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. Trans Assoc Comput Linguist 2017 Dec;5:135-146 [FREE Full text] [doi: 10.1162/tacl_a_00051]
12. Peters M, Neumann M, Zettlemoyer L, Yih W. Dissecting contextual word embeddings: Architecture and representation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium p. 1499-1509 URL: https://www.aclweb.org/anthology/D18-1179.pdf [doi: 10.18653/v1/D18-1179]
13. Radford A, Narasimhan K, Salimans T, Sutskever I. OpenAI preprint. 2018. Improving language understanding by generative pre-training URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed 2020-04-12]
14. Devlin J, Chang MW, Lee K, Toutanova K. arXiv. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding URL: https://arxiv.org/pdf/1810.04805.pdf [accessed 2019-10-18]
15. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee L, et al. Deep contextualized word representations. In: Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). 2018 Presented at: 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018); June 1-6, 20188; New Orleans, LA p. 2227-2237 URL: https://www.aclweb.org/anthology/N18-1202.pdf [doi: 10.18653/v1/n18-1202]
16. Stubbs A, Uzuner Ö. Annotating risk factors for heart disease in clinical narratives for diabetic patients. J Biomed Inform 2015 Dec;58 Suppl:S78-S91 [FREE Full text] [doi: 10.1016/j.jbi.2015.05.009] [Medline: 26004790]
17. Huang Z, Xu W, Yu K. arXiv. 2015. Bidirectional LSTM-CRF models for sequence tagging URL: https://arxiv.org/pdf/1508.01991v1.pdf [accessed 2019-10-18]
18. Cai X, Dong S, Hu J. A deep learning model incorporating part of speech and self-matching attention for named entity recognition of Chinese electronic medical records. BMC Med Inform Decis Mak 2019 Apr 09;19(Suppl 2):65 [FREE Full text] [doi: 10.1186/s12911-019-0762-7] [Medline: 30961622]
19. Wu Y, Jiang M, Lei J, Xu H. Named entity recognition in Chinese clinical text using deep neural network. Stud Health Technol Inform 2015;216:624-628 [FREE Full text] [Medline: 26262126]
20. Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition. In: Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018). 2018 Presented at: 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018); June 1-6, 2018; New Orleans, LA p. 1446-1459 URL: https://www.aclweb.org/anthology/N18-1131.pdf [doi: 10.18653/v1/N18-1131]
21. Rei M, Crichton GKO, Pyysalo S. Attending to characters in neural sequence labeling models. In: Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016): Technical Papers. 2016 Presented at: 26th

International Conference on Computational Linguistics (COLING 2016): Technical Papers; December 11-17, 2016; Osaka, Japan p. 309-318 URL: https://www.aclweb.org/anthology/C16-1030.pdf
22. Lee JY, Dernoncourt F, Szolovits P. arXiv. 2017. Transfer learning for named-entity recognition with neural networks URL: https://arxiv.org/pdf/1705.06273.pdf [accessed 2019-10-18]
23. Peng D, Wang Y, Liu C, Chen Z. TL-NER: A transfer learning model for Chinese named entity recognition. Inf Syst Front 2019 Jun 4:1. [doi: 10.1007/s10796-019-09932-y]
24. Bharadwaj A, Mortensen D, Dyer C, Carbonell J. Phonologically aware neural model for named entity recognition in low resource transfer settings. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016 Presented at: 2016 Conference on Empirical Methods in Natural Language Processing; November 1-5, 2016; Austin, TX p. 1462-1472 URL: https://www.aclweb.org/anthology/D16-1153.pdf [doi: 10.18653/v1/D16-1153]
25. Takase S, Suzuki J, Nagata M. Direct output connection for a high-rank language model. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018 Presented at: 2018 Conference on Empirical Methods in Natural Language Processing; October 31-November 4, 2018; Brussels, Belgium p. 4599-4609 URL: https://www.aclweb.org/anthology/D18-1489.pdf [doi: 10.18653/v1/D18-1489]
26. Kaneko M, Komachi M. arXiv. 2019. Multi-head multi-layer attention to deep language representations for grammatical error detection URL: https://arxiv.org/pdf/1904.07334.pdf [accessed 2019-10-18]

Abbreviations

Abn: abnormal inspection result
BERT: Bidirectional Encoder Representations from Transformers
BiLSTM: bidirectional long short-term memory
BIO: beginning of the noun phrase, middle of the noun phrase, and not a noun phrase
CCKS: China Conference on Knowledge Graph and Semantic Computing
CEMR: Chinese electronic medical record
CNN: convolutional neural network
CRF: conditional random field
Dis: disease
ELMo: Embeddings from Language Models
EMR: electronic medical record
ER: entity recognition
F1: F1 score
GloVe: Global Vectors
GPT: Generative Pretraining Transformer
K: key
LSTM: long short-term memory
Med: medicine
multi-level ER: multi-level representation learning for entity recognition
NLP: natural language processing
P: precision
Q: query
R: recall
RNN: recurrent neural network
SM: self-matching attention mechanism
Sym: symptom
Tre: treatment
V: value

Edited by T Hao; submitted 30.12.19; peer-reviewed by W Song, L Li; comments to author 14.02.20; revised version received 24.02.20; accepted 19.03.20; published 04.05.20.

Please cite as:
Zhang Z, Zhu L, Yu P
Multi-Level Representation Learning for Chinese Medical Entity Recognition: Model Development and Validation
JMIR Med Inform 2020;8(5):e17637
URL: https://medinform.jmir.org/2020/5/e17637
doi:10.2196/17637
PMID:32364514

Original Paper

A Graph Convolutional Network–Based Method for Chemical-Protein Interaction Extraction: Algorithm Development

Erniu Wang[1*], MS; Fan Wang[1*], PhD; Zhihao Yang[1*], PhD; Lei Wang[2*], PhD; Yin Zhang[2*], PhD; Hongfei Lin[1*], PhD; Jian Wang[1*], PhD

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, China

[2]Beijing Institute of Health Administration and Medical Information, Beijing, China

[*]all authors contributed equally

Corresponding Author:
Zhihao Yang, PhD
College of Computer Science and Technology
Dalian University of Technology
No 2, Linggong Road
Ganjingzi District
Dalian,
China
Phone: 86 13190114398
Email: yangzh@dlut.edu.cn

Abstract

Background: Extracting the interactions between chemicals and proteins from the biomedical literature is important for many biomedical tasks such as drug discovery, medicine precision, and knowledge graph construction. Several computational methods have been proposed for automatic chemical-protein interaction (CPI) extraction. However, the majority of these proposed models cannot effectively learn semantic and syntactic information from complex sentences in biomedical texts.

Objective: To relieve this problem, we propose a method to effectively encode syntactic information from long text for CPI extraction.

Methods: Since syntactic information can be captured from dependency graphs, graph convolutional networks (GCNs) have recently drawn increasing attention in natural language processing. To investigate the performance of a GCN on CPI extraction, this paper proposes a novel GCN-based model. The model can effectively capture sequential information and long-range syntactic relations between words by using the dependency structure of input sentences.

Results: We evaluated our model on the ChemProt corpus released by BioCreative VI; it achieved an F-score of 65.17%, which is 1.07% higher than that of the state-of-the-art system proposed by Peng et al. As indicated by the significance test (P<.001), the improvement is significant. It indicates that our model is effective in extracting CPIs. The GCN-based model can better capture the semantic and syntactic information of the sentence compared to other models, therefore alleviating the problems associated with the complexity of biomedical literature.

Conclusions: Our model can obtain more information from the dependency graph than previously proposed models. Experimental results suggest that it is competitive to state-of-the-art methods and significantly outperforms other methods on the ChemProt corpus, which is the benchmark data set for CPI extraction.

(JMIR Med Inform 2020;8(5):e17643) doi:10.2196/17643

KEYWORDS

chemical-protein interaction; graph convolutional network; long-range syntactic; dependency structure

Introduction

Biomedical literature has grown significantly with the development of biomedical technology, which contains a large amount of valuable chemical-protein interactions (CPIs). CPI extraction plays an important role in various biomedical tasks such as drug discovery, medicine precision, and knowledge graph construction [1]. With the rapidly increasing volume of biomedical literature, it becomes time-and-resource–consuming to extract CPIs from biomedical literature manually. There are

some computational methods that have been successfully proposed for automatic biomedical relation extraction [2-6]. However, most previous studies focused on the extraction of drug-drug interactions, protein-protein interactions, and chemical-disease interactions; only a few attempts were developed to extract CPIs [7].

The BioCreative VI ChemProt shared task [8] created the ChemProt data set, which is used in the development of CPI extraction methods. The current CPI extraction systems can be generally divided into two categories: the traditional machine learning–based methods and the neural network–based methods. The traditional machine learning–based methods conventionally train a CPI extractor by handcrafted features [7]. The neural network–based methods can automatically learn powerful features to train a classifier, and therefore, have become a promising method for CPI extraction.

Mehryary et al [9] combined a support vector machine (SVM) and long short-term memory (LSTM) to extract CPIs and achieved a high F-score by a rich set of features. Warikoo et al [10] also exploited a set of linguistic features to train a tree kernel classifier to obtain CPIs from biomedical literature. Generally, these methods depend heavily on feature engineering. Recently, attention mechanisms have been successfully used in many natural language processing tasks, and some works have employed it in CPI extraction. Liu et al [11] aggregated an attention mechanism and gated recurrent units (GRU) to extend the LSTM model. Verga et al [12] encoded pair-wise predictions over entire abstracts by synthesizing self-attention and convolutions. Corbett and Boyle [13] employed multiple LSTM layers with unlabeled data to extract relations amongst the ChemProt corpus and achieved good performance. Peng et al [14] applied an ensemble system to extract CPIs, which consists of three individual models, including SVM, convolutional neural network (CNN), and bi-directional long short-term memory (Bi-LSTM) modules. The system achieved an F-score of 64.1% and won the top rank in the BioCreative VI ChemProt shared task.

However, most of the proposed methods only utilize the sequential information of sentences; syntactic information has not been carefully studied yet. Due to the presence of complex sentences in biomedical literature, it is difficult to effectively learn the semantic and syntactic information for some neural network–based models (eg, CNN [15], LSTM [13,16], and GRU [17]). To address this problem, we apply a graph convolutional network (GCN) [18,19] for CPI extraction. The GCN can exploit dependency structure and capture long-range syntactic relations of input sentences. Therefore, it is more effective and precise than other modules for CPI extraction.

Additionally, sentences in the biomedical literature are generally lengthy, so there is a considerable amount of irrelevant words. For example, in the sentence “Dasatinib (BMS-354825) is a novel orally bioavailable SRC/ABL inhibitor that has activity against multiple imatinib-resistant BCR-ABL isoforms in vitro that is presently showing considerable promise in early-phase clinical trials of chronic myeloid leukemia (CML),” “Dasatinib (BMS-354825) is a novel orally bioavailable SRC/ABL inhibitor” can already express the inhibitory relationship between the entities “Dasatinib” and “SRC.” Other words, which may affect the performance of the relation extractor, are irrelevant. Inspired by Zhang et al [20], we apply a path-centric pruning strategy to incorporate relevant information while maximally reducing the influence of noisy words in long sentences. This strategy retains tokens that are up to distance N away from the dependency path in the lowest common ancestor (LCA) subtree [21]. The experimental results prove that this strategy can improve the robustness of our model. The model achieves the best balance between noisy words and relevant words when N is set to 2.

A single GCN model usually depends highly on correct parse trees to extract crucial information from sentences, while existing parsing algorithms produce imperfect trees in many cases. To further improve the robustness of our mode, we apply a Bi-LSTM network to obtain contextual information about word order or disambiguation. The compound model can better leverage local word patterns regardless of parsing quality.

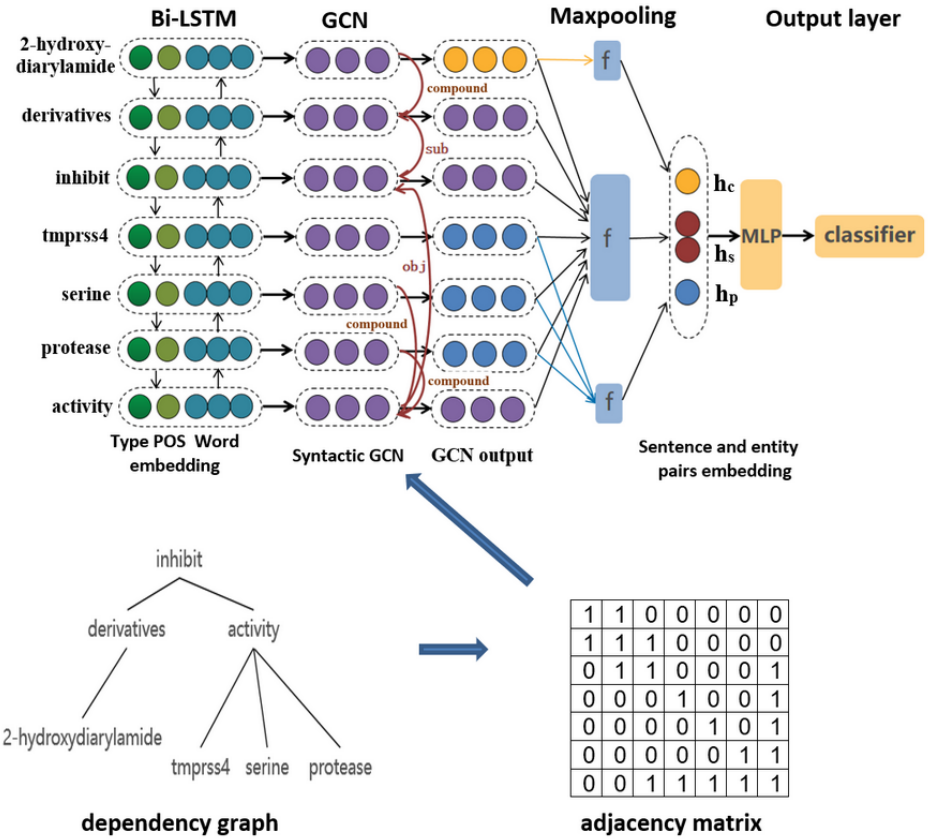
In summary, we propose a GCN-based model in this paper to extract CPIs. We evaluated our model on the ChemProt corpus, which is the benchmark data set for CPI extraction. To the best of our knowledge, this is the first study to use a GCN encoding syntactic graph for CPI extraction.

Methods

Overview

The overall architecture of our model is presented in Figure 1. Our model contains three parts: the Bi-LSTM layer, the GCN layer, and the classification layer. In the model, a Bi-LSTM layer is applied first to capture local word patterns and output the representation of each word within the whole sentence. Subsequently, the contextualized representation and the dependency graph (with two directly attached dependencies) of input sentences are fed into the GCN layer to integrate dependency information into word representations. After that, a max-pooling layer is applied to generate the representation of the sentence and two target entities from word representations. Finally, these representations are concatenated and fed into a multilayer perceptron (MLP) for softmax classification. In the following section, we will introduce our model in detail.

Figure 1. The overall architecture of our model. Bi-LSTM: bi-directional long short-term memory; GCN: graph convolutional network; POS: part-of-speech; MLP: multilayer perceptron; sub: subject; obj: object; hc: representation of chemical; hs: representation of sentence; hp: representation of protein; f: max-pooling function.



The Bi-LSTM Layer

We adopt a Bi-LSTM layer to capture contextual information about word order and reduce the impact of parsing errors in our model. The Bi-LSTM layer is applied on the whole sentence to learn the representation of each word. Bi-LSTM can capture more comprehensive features by dealing with the input sequence from forward and backward directions, compared with unidirectional LSTM; it is the combination of the forward LSTM and backward LSTM.

In the ChemProt corpus, some entities contain multiple types of words, especially the relation type "PART_OF," which means one entity is part of another type of entity within a relation entity pair. For example, "thiazide-sensitive sodium-chloride cotransporter" is a gene entity, and "sodium-chloride" is a chemical entity. To reduce this interference, we apply prior knowledge of the entity type as a feature to improve CPI extraction.

The input of the Bi-LSTM layer consists of three parts, including word embedding, part-of-speech (POS) embedding and entity type embedding. Given a sentence $S = \{w_1,w_2,\ldots,w_n\}$, the POS sequence $P = \{p_1,p_2,\ldots,p_n\}$ can be obtained by the Stanford CoreNLP toolkit [22], where w_i is the i-th word in a sentence and p_i is its POS. We obtain the sequence of entity types $T = \{t_1,t_2,\ldots,t_n\}$ through the index information of the entity pairs in a sentence. We tagged entity tokens "chemical" or "gen" and other words "O." The word embedding is initialized with pretrained word embedding, which is obtained by FastText [23]. POS and entity type embedding are initialized randomly. The input of the model is denoted as follows:

For each token x_i, the forward LSTM and backward LSTM consider the contextual information before and after it, respectively. The final output is the concatenation of the two directions. The Bi-LSTM calculation process is presented as follows:

where  and  denote the hidden states of the forward and backward LSTM of x_i, respectively.  denotes concatenation operation.

The GCN Layer

GCNs can learn a state embedding, which contains the information of a neighborhood for each node in a graph. It has been proven that models or dependency-based models are very effective in relation extraction by capturing long-range syntactic relations [24-26]. In our model, we apply a GCN to improve the performance of CPI extraction by utilizing the dependency parse trees of the input sentences. In order to reduce the influence of noisy words in long sentences, we further apply a pruning strategy on the dependency trees to remove irrelevant words while maximally keeping crucial content.

Given a sentence, we first apply the Stanford CoreNLP toolkit to get its dependency tree, which is considered as an undirected graph. Then, we apply a path-centric pruning strategy and retain two directly attached words around the shortest path at the LCA of the two entities [20]. After that, we convert the subgraph into an adjacency matrix A. If there is a dependency relation between node *i* and *j*, is assigned with a value of 1. Finally, we apply a GCN over the output of Bi-LSTM and adjacency matrix A to get an updated hidden representation of h_i. This can be represented as shown in formula 5. In an L-layer GCN, if we use  as the input vector and  as the output vector for node *i* at the l-th layer, the graph convolution operation of the l-th layer can be represented as shown in formula 6.

where $W^{(j)}$ and $W^{(l)}$ are weight linear transformations, $b^{(j)}$ and $b^{(l)}$ are bias terms, and *f* is a nonlinear function (eg, a rectified linear unit [ReLU]). We could obtain the hidden representation of each token directly influenced by its neighbors no more than *L* edges apart in the dependency trees after applying an L-layer GCN over word vectors. To avoid a sentence representation favoring high-degree nodes regardless of the information carried in the node and to transfer information in  to  , we normalized the activations in the graph convolution before feeding it through a nonlinearity, and added self-loops to each node in the graph:

where  . *I* is the $n \times n$ identity matrix, and  is the degree of token *i* in the resulting graph.

The Output and Classification Layer

The CPI extraction can be regarded as a classification problem. Given a sentence $S = \{w_1,w_2,...,w_n\}$ where w_i is the *i*-th token, let $S_c = \{w_{c1},w_{c2},...,w_{cn}\}$ and $S_p = \{w_{p1},w_{p2},...,w_{pn}\}$ denote chemical sequence and protein sequence, respectively. The goal of CPI extraction is to predict the relation rR hold between the chemical S_c and gen S_p; otherwise, "no relation" is declared. After the Bi-LSTM and GCN layers, we can obtain the hidden representation of each token, which is influenced by not only local word patterns but also long-range words. To utilize these word representations for relation extraction, we mapped from $h^{(L)}$ (*n* output vectors) to the sentence vector h_{sent}. The information close to entity tokens in the dependency trees is generally important in relation classification. Therefore, we also apply a max-pooling function to obtain entity pair representations h_c and h_p from $h^{(L)}$ as follows:

where  denotes the output after L-layer GCN, and *f* denotes a max-pooling function.

Then, we connect sentence representation with entity representation [27,28] as a new representation, and feed it into a feed-forward neural network (FFNN) inspired by relational reasoning works:

Finally, we apply a linear layer followed by a softmax operation over the final representation h_{final} to obtain a probability distribution over chemical-protein relations and the computation is shown as follows:

where W_r and b_r are trainable parameters, and *r* is relation type.

Evaluation Metrics

In experiments, the Micro-average F-score is applied to evaluate the performance of our model, which is a harmonic mean of *P* and *R*, where *P* denotes precision and *R* denotes recall:

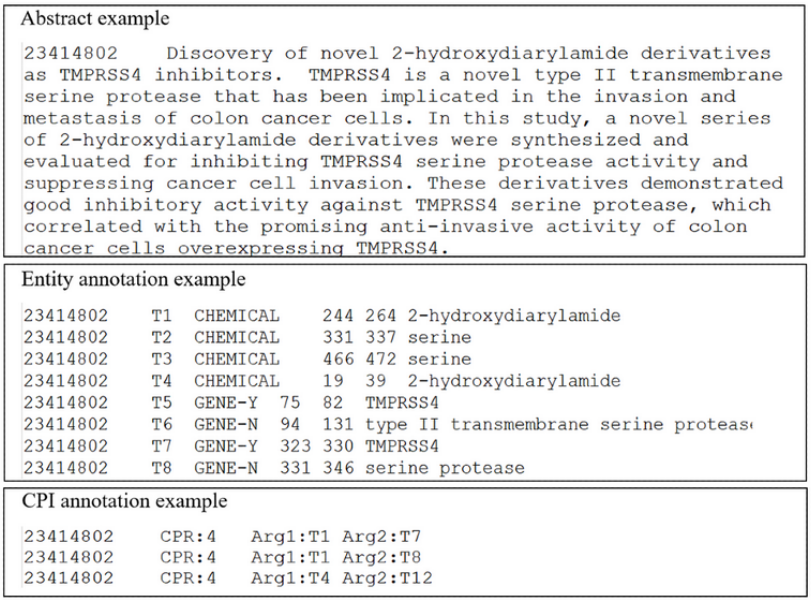
TP, *FN*, and *FP* denote true positive, false negative, and false positive, respectively.

Results

Data Retrieval and Preprocessing

CPI extraction aims to classify whether a semantic relation that holds between the chemical and protein entity pairs within a sentence or document. The BioCreative VI ChemProt task delivered the corpus as a manually annotated CPI data set that consists of training, development, and test sets. Each set includes the abstracts, entities, and relations files. Figure 2 provides an example of the three files from the ChemProt training set.

Figure 2. Examples of the ChemProt corpus. CPI: chemical-protein interaction.



The abstracts file provides the article identifier, title, and abstract document for each article. The entities file consists of the PubMed Unique Identifier (PMID), entity number, type of entity mentions, start and end character offset, and text string of entity mention. The relations file is composed of the PMID, CPI relation class, evaluation type, and CPI relation and interactor arguments. In the ChemProt corpus, there are 10-type relation classes, and each relation class includes one or multiple relation types (Table 1). Although there are 10-type relation classes in ChemProt corpus, only five are used for evaluation purposes (ie, CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9). Table 2 shows the statistics of the ChemProt corpus.

The original corpus consists of PubMed abstracts from biomedical literature in which more than 98% of relation entity pairs within a sentence [8]. Therefore, we neglected the cross-sentence entity pairs and conducted experiment at the sentence level. For CPI extraction, we took some preprocessing steps on the original corpus. First, we split abstracts into sentences and only retained the sentences that contained the relational entity pairs. Then, we reassigned the training set and developing set with a ratio of 9:1. Finally, we replaced each digit string that was not an entity substring with a particular "num" tag.

Figure 3 gives two illustrative examples of CPI extraction. In the first example, the sentence "Alprenolol and BAAM also caused surmountable antagonism of isoprenaline responses, and this beta 1-adrenoceptor antagonism was slowly reversible." contains a relational entity pair. To accurately extract the CPI, we need to first detect the chemical entity "Alprenolol" and protein entity "beta 1-adrenoceptor," and then classify the interaction as the CPR:6 class. The second example is a long and complex sentence. It is more difficult for the relation classifier to extract the interaction between the chemical and protein entities. Our model aims to predict the interactions, and the output is the relation type of chemical-protein entity pairs as shown in Figure 3.

Table 1. The chemical-protein relation (CPR) groups.

Group	Evaluated in the BioCreative VI ChemProt shared task?	ChemProt relations
CPR:1	No	PART_OF
CPR:2	No	REGULATOR|DIRECT_REGULATOR|INDIRECT_REGULATOR
CPR:3	Yes	UPREGULATOR|ACTIVATOR|INDIRECT_UPREGULATOR
CPR:4	Yes	DOWNREGULATOR|INHIBITOR|INDIRECT_DOWNREGULATOR
CPR:5	Yes	AGONIST|AGONIST-ACTIVATOR|AGONIST-INHIBITOR
CPR:6	Yes	ANTAGONIST
CPR:7	No	MODULATOR|MODULATOR-ACTIVATOR|MODULATOR-INHIBITOR
CPR:8	No	COFACTOR
CPR:9	Yes	SUBSTRATE|PRODUCT_OF|SUBSTRATE_PRODUCT_OF
CPR:10	No	NOT

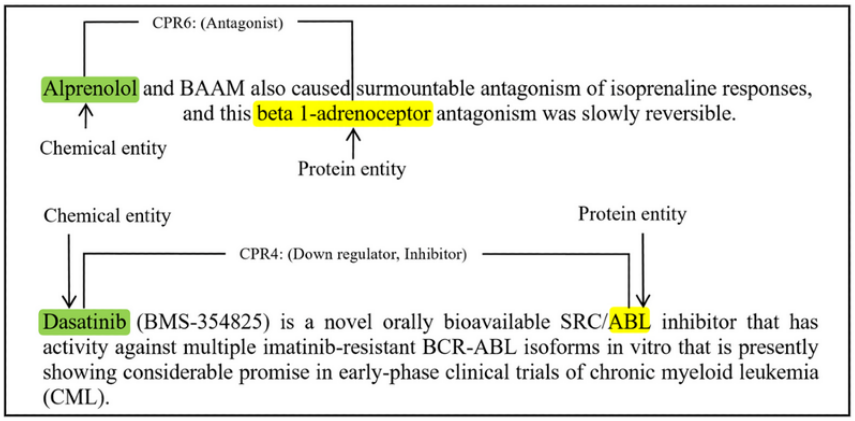
Table 2. Statistics of the ChemProt corpus.

Annotations	Data set		
	Training, n	Development, n	Test, n
Document	1020	612	800
Chemicals	13,017	8004	10,810
Proteins	12,752	7567	10,019
CPR[a]:3	768	550	665
CPR:4	2254	1094	1661
CPR:5	173	116	195
CPR:6	235	199	293
CPR:9	727	457	644
Evaluated CPIs[b]	4157	2416	3458
Evaluated CPIs in one sentence	4122	2412	3444

[a]CPR: chemical-protein relation.

[b]CPI: chemical-protein interaction.

Figure 3. Illustrative examples of chemical-protein relation (CPR) classes.



Experimental Settings

In this work, FastText [23] was used to pretrain word embedding on the ChemProt corpus. Before the experiments, we set the range of parameters based on experience, then tuned the parameters on the development set by using grid search to determine the optimal parameters, and finally selected the best model of parameters that were optimal for evaluation on the test set. Without overfitting, the best model generally can achieve the best performance (the highest F-score) on the development set. The detailed tune range and hyperparameter values are listed in Table 3.

Comparison of Different Pruning Distances

To obtain the best pruning distance, we experimented with $N\{0,1,2,3,\infty\}$ on the ChemProt corpus—N=0 corresponds to pruning the tree down to the path; N=1 keeps all nodes that are directly attached to the path; N=2,3 means holding words up to distance 2 and 3 away from the dependency path in the LCA subtree; and N=∞ retains the entire LCA subtree.

As shown in Figure 4, the performance of our model reaches its peak and outperforms other pruning distance at N=2. This confirms that pruning too aggressively (N=0,1) could lead to a loss of crucial information while retaining too many irrelevant words (N=3) also decreases model performance due to the interference of irrelevant information. When N=2, the model achieves the best balance between including relevant and irrelevant information.

Table 3. Hyperparameter setting.

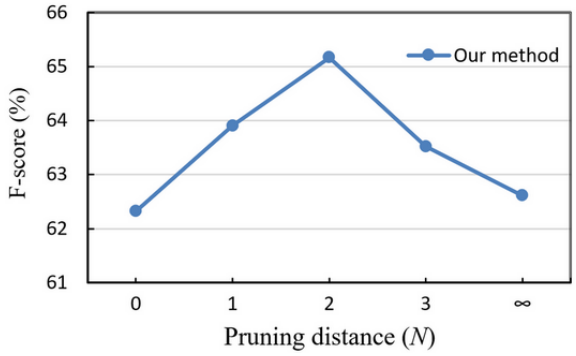
Hyperparameter	Tuned range	Optimal
Word embedding dimension	[100,200,300]	200
POS[a] embedding dimension	[10,20,30,40]	20
Entity type embedding dimension	[40,50,60,70,80]	60
GCN[b] hidden units	[100,200,300]	200
LSTM[c] hidden units	[100,200,300]	200
Learning rate	[0.1,0.2,0.3,0.4]	0.3
Dropout rate	[0.4,0.5,0.6]	0.5

[a]POS: part-of-speech.

[b]GCN: graph convolutional network.

[c]LSTM: long short-term memory.

Figure 4. Comparison of different pruning distances.



Comparison of Different Embedding Features

Table 4 shows the effectiveness of different embedding features, including word embedding, entity type embedding, and POS embedding. The model achieves an F-score of 59.56% when only using word embedding. When POS and word embedding are combined, the F-score increases to 60.69%. When the entity type and word embedding are combined, the F-score increases to 62.52% (an increase of 2.96%). Furthermore, when both entity type and POS embedding are integrated with word embedding, the F-score improves to 65.17%. The results suggest that the main contributor to performance is prior knowledge of the entity type. This confirms the validity of the entity type in CPI extraction. The POS embedding is also valuable to the model.

Table 4. Performance evaluation of different embedding features.

Embedding feature	Precision (%)	Recall (%)	F-score (%)	Δ (%)
Word	57.64	61.62	59.56	—[a]
Word+POS[b]	58.49	63.06	60.69	+1.13
Word+Entity type	64.06	61.05	62.52	+2.96
Word+POS+Entity type	63.79	66.62	65.17	+5.61

[a]Not applicable.

[b]POS: part-of-speech.

Comparison With the Baseline Method

Different single models and their ensemble models are compared with each other in this section. As shown in Table 5, all ensemble models perform better than all single models, and the GCN+Bi-LSTM model performs better than the Bi-LSTM+CNN model. The results indicate that ensemble models can generally capture more information than single models. In terms of overall performance, the precision, recall, and F-score of the Bi-LSTM+GCN model are higher than those of the Bi-LSTM+CNN model. Our model can fully capture the overall information of the sentence by combining sequence structure information and syntactic information, while the Bi-LSTM+CNN model could only obtain sequence structure information, which confirms the effectiveness of the GCN model in CPI extraction.

Table 5. Comparison with the baseline method.

Model	Precision (%)	Recall (%)	F-score (%)
Single models			
CNN[a]	42.47	69.43	52.70
GCN[b]	48.77	63.69	55.24
Bi-LSTM[c]	60.59	60.34	60.46
Ensemble models			
Bi-LSTM+CNN	57.77	64.73	61.05
Bi-LSTM+GCN (our model)	63.79	66.62	65.17

[a]CNN: convolutional neural network.

[b]GCN: graph convolutional network.

[c]Bi-LSTM: bi-directional long short-term memory.

Discussion

The experimental results suggest that our model can effectively extract CPIs; it is better at learning semantic and syntactic information from sentences compared to other models. Additionally, the pruning strategy can alleviate the influence of irrelevant words in long sentences in biomedical literature, by only retaining *N* away tokens from the dependency path in the LCA subtree.

Comparison With Prior Work

A comparison of our model with other existing methods on the ChemProt corpus is shown in Table 6. It can be found that neural network–based methods perform better than traditional machine learning–based methods, and our method achieves the highest F-score of 65.17%.

Table 6. Comparison with other existing methods.

Model	Precision (%)	Recall (%)	F-score (%)
Verga et al [12]	48.00	54.10	50.80
Matos [29]	57.38	47.22	51.81
Liu et al [11]	57.4	48.7	52.7
Lung et al [30]	63.52	51.21	56.71
Corbett and Boyle [13]	62.97	62.20	62.58
Mehryary et al [9]	59.05	67.76	63.10
Peng et al [14]	72.66	57.35	64.10
Our model	63.79	66.62	65.17

Lung et al [30] used machine learning methods to integrate the semantic and dependency graph features through a three-stage model. They achieved an F-score of 56.71%. Similarly, Corbett and Boyle [13] used pretrained LSTM and Bi-LSTM to extract CPIs in two stages and achieved a higher F-score of 61.5%. A particular feature of their system was the usage of unlabeled data both to pretrain word embedding and pretrain LSTM layers in the neural network.

Verga et al [12] applied attention mechanisms in their model. They synthesized convolutions and self-attention to extract CPIs. Liu et al [11] achieved an F-score of 52.7% by synthesizing GRU and attention pooling. The results of word-level attention weights in the model of Liu et al [11] showed that attention mechanism is effective in selecting the most important trigger words when trained with semantic relation labels without the need of semantic parsing and feature engineering.

Mehryary et al [9] employed an ensemble system that combined the results of SVM and LSTM, and they achieved a competitive result. Peng et al [14] utilized more external features. They stacked SVM, CNN, and RNN models, and combined the outputs of the three systems by either majority voting or stacking. They achieved the best F-score of 64.10% in the BioCreative VI ChemProt shared task. Our model synthesized Bi-LSTM and GCN and achieved an improvement of 1.07% in F-score over the system of Peng et al [14]. We further performed significance tests with $P<.05$ indicating significance. The P value of Peng et al [14] and our model is less than .001. It indicates that the improvement of 1.07% in F-score is significant.

Results Analysis

The experimental results indicate that the GCN module is valuable in CPI extraction. It can extract CPIs from biomedical texts with syntactic graph representations. It might be also efficient in other biomedical tasks by utilizing the sentence

parse structure. By comparing different pruning distance, we revealed that the length of sentence also plays an important role in relation extraction. The noisy words that are irrelevant to relations might hamper the performance of the extractor.

GCNs can learn effective representation for relation extraction. However, a single GCN model could not capture the contextual information of word order. Additionally, GCN highly depends on correct parse trees to extract information from sentences, while existing parsing algorithms produce imperfect trees in many cases. To resolve these issues and improve the robustness of our model, we applied Bi-LSTM to generate contextualized representation and feed it into the GCN layer. The results confirm that the ensemble model of GCN and Bi-LSTM is validated for CPI extraction.

Contributions

The model we proposed in this paper aims to extract CPI and achieve state-of-the-art performance on the ChemProt corpus. Our main contributions are as follows.

We proposed a novel neural model based on a GCN for CPI extraction, which can capture long-range syntactic information by utilizing the dependency structure of the input sentence. To improve the robustness, we applied a path-centric pruning strategy to remove irrelevant words without damaging crucial content on the dependency trees. Through the pruning strategy, the influence of noisy words can be reduced, thereby further improving the performance of the model. Furthermore, a Bi-LSTM layer is utilized to better leverage local word patterns regardless of parsing quality.

Our model can automatically extract CPIs from a large amount of biomedical literature, which can save significant labor force and resources. Abundant biological entity relations can deliver useful chemicals for some diseases and save time by optimizing the drug development cycle, thereby helping pharmacists discover drugs. Furthermore, the knowledge graph generally contains rich, structured knowledge and has been widely used in natural language processing applications, such as search engines and question answering systems. However, the rapidly increasing volume of information requires refinement in the coverage of knowledge graphs. CPI extraction can help researchers to efficiently acquire biomedical knowledge, which can enrich the information needed for knowledge graph construction.

Conclusions

We proposed a novel model based on a GCN to extract CPI. The GCN module can encode syntactic information over the dependency graphs of input sentences. To reduce the impact of noisy words, our model only retains tokens that are up to a distance of N=2 away from the dependency path in the LCA subtree. Additionally, it applies Bi-LSTM to generate a contextualized representation and feed it into the GCN layer to resolve parsing errors and improve the robustness of the model. The experimental results demonstrated that our model achieves state-of-the-art performance. We plan to further improve our model and apply our method to extract other biomedical relation entity pairs.

Acknowledgments

We appreciate the valuable feedback provided by three anonymous reviewers. EW carried out the overall algorithm design and experiments as well as the writing of the manuscript. FW, ZY, LW, YZ, HL, and JW contributed to the algorithm design and the writing of the manuscript. All authors read and approved the final manuscript. This work was supported by a grant from the National Key Research and Development Program of China (#2016YFC0901902).

Conflicts of Interest

None declared.

References

1. Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI, Taboureau O. ChemProt-3.0: a global chemical biology diseases mapping. Database (Oxford) 2016 Feb 13;2016 [FREE Full text] [doi: 10.1093/database/bav123] [Medline: 26876982]
2. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics 2008 Nov 19;9 Suppl 11(S11):S2 [FREE Full text] [doi: 10.1186/1471-2105-9-S11-S2] [Medline: 19025688]
3. Miwa M, Sætre R, Miyao Y, Tsujii J. A rich feature vector for protein-protein interaction extraction from multiple corpora. In: Association for Computational Linguistics. 2009 Presented at: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1; 2009; Tokyo p. 121-130. [doi: 10.3115/1699510.1699527]
4. Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics 2010 Feb 25;11(1). [doi: 10.1186/1471-2105-11-107]
5. Zhang Y, Lin H, Yang Z, Wang J, Li Y. A Single Kernel-Based Approach to Extract Drug-Drug Interactions from Biomedical Literature. PLoS ONE 2012 Nov 1;7(11):e48901. [doi: 10.1371/journal.pone.0048901]
6. Segura-Bedmar I, Martínez P, Herrero-Zazo M. Lessons learnt from the DDIExtraction-2013 Shared Task. Journal of Biomedical Informatics 2014 Oct;51:152-164. [doi: 10.1016/j.jbi.2014.05.007]
7. Krallinger M, Rabal O, Lourenço A, Oyarzabal J, Valencia A. Information Retrieval and Text Mining Technologies for Chemistry. Chem. Rev 2017 May 05;117(12):7673-7761. [doi: 10.1021/acs.chemrev.6b00851]

8. Krallinger M, Rabal O, Akhondi S. Overview of the BioCreative VI chemical-protein interaction track. 2017 Presented at: Proceedings of the sixth BioCreative challenge evaluation workshop; 2017; Bethesda, MD, USA p. 141-146.
9. Mehryary F, Björne J, Salakoski T, Ginter F. Potent pairing: ensemble of long short-term memory networks and support vector machine for chemical-protein relation extraction. Database (Oxford) 2018 Jan 01;2018:1-23 [FREE Full text] [doi: 10.1093/database/bay120] [Medline: 30576487]
10. Warikoo N, Chang Y, Hsu W. LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task. Database (Oxford) 2018 Jan 01;2018:1-21 [FREE Full text] [doi: 10.1093/database/bay108] [Medline: 30346607]
11. Liu S, Shen F, Komandur Elayavilli R, Wang Y, Rastegar-Mojarad M, Chaudhary V, et al. Extracting chemical-protein relations using attention-based neural networks. Database (Oxford) 2018 Jan 12;2018:1-12 [FREE Full text] [doi: 10.1093/database/bay102] [Medline: 30295724]
12. Verga P, Strubell E, Mccallum A. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. 2018 Presented at: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers); 2018; New Orleans, Louisiana p. 872-884.
13. Corbett P, Boyle J. Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings. Database (Oxford) 2018 Jan 10;2018:1-10 [FREE Full text] [doi: 10.1093/database/bay066] [Medline: 30010749]
14. Peng Y, Rios A, Kavuluru R, Lu Z. Extracting chemical-protein relations with ensembles of SVM and deep learning models. Database (Oxford) 2018 Jan 09;2018:1-9 [FREE Full text] [doi: 10.1093/database/bay073] [Medline: 30020437]
15. Yüksel A, Öztürk H, Ozkirimli E, Özgür A. CNN-based chemical-protein interactions classification. 2017 Presented at: Proceedings of the BioCreative VI Workshop; 2017; Boğaziçi University, İstanbul, Turkey p. 184-186.
16. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation 1997 Nov;9(8):1735-1780. [doi: 10.1162/neco.1997.9.8.1735]
17. Cho K, Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder? Decoder for statistical machine translation (arXiv:1406.1078). arXiv.org 2014:1724-1734 [FREE Full text]
18. Kipf T, Welling M. Semi-Supervised Classification with Graph Convolutional Networks (arXiv:1609.02907). arXiv.org 2017 Feb 22:1-14 [FREE Full text]
19. Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling (arXiv:1703.04826). arXiv.org 2017:1-11 [FREE Full text]
20. Zhang Y, Qi P, Manning CD. Graph Convolution over Pruned Dependency Trees Improves Relation Extraction. 2018 Presented at: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium p. 2205-2215 URL: https://dblp.uni-trier.de/db/conf/emnlp/emnlp2018.html#Zhang0M18
21. Miwa M, Bansal M. End-to-end relation extraction using LSTMs on sequences and tree structures (arXiv:1601.00770). arXiv preprint 2016:1-13 [FREE Full text]
22. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. 2014 Presented at: Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations; June 23-24, 2014; Baltimore, Maryland USA p. 55-60.
23. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 2017;5:135-146. [doi: 10.1162/tacl_a_00051]
24. Guo Z, Zhang Y, Lu W. Attention Guided Graph Convolutional Networks for Relation Extraction. 2019 Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy p. 241-251 URL: https://dblp.uni-trier.de/db/conf/acl/acl2019-1.html#GuoZL19
25. Sahu SK, Christopoulou F, Miwa M, Ananiadou S. Inter-sentence Relation Extraction with Document-level Graph Convolutional Neural Network (arXiv:1906.04684). arXiv.org 2019 [FREE Full text]
26. Zhang N, Deng S, Sun Z, Wang G, Chen X, Zhang W, et al. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks (arXiv:1903.01306). arXiv.org 2019 [FREE Full text]
27. Lee K, He L, Lewis M, Zettlemoyer L. End-to-end neural coreference resolution (arXiv:1707.07045). arXiv.org 2017 [FREE Full text]
28. Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, et al. A simple neural network module for relational reasoning. 2017 Presented at: Advances in neural information processing systems; 2017; Long Beach, CA, USA.
29. Matos S. Extracting chemical–protein interactions using long short-term memory networks. 2017 Presented at: Proceedings of the BioCreative VI Workshop; 18-20 October 2017; Bethesda.
30. Lung PY, Zhao T, He Z. Extracting chemical-protein interactions from literature. 2017 Presented at: Proceedings of the BioCreative VI Workshop; 2017; Florida State University, Tallahassee, FL, 32306 USA.

Abbreviations

Bi-LSTM: bi-directional long short-term memory
CNN: convolutional neural network

CPI: chemical-protein interaction
CPR: chemical-protein relation
FFNN: feed-forward neural network
FN: false negative
FP: false positive
GCN: graph convolutional network
GRU: gated recurrent units
LCA: lowest common ancestor
LSTM: long short-term memory
MLP: multilayer perceptron
PMID: PubMed Unique Identifier
POS: part-of-speech
ReLU: rectified linear unit
SVM: support vector machine
TP: true positive

Edited by T Hao, B Tang; submitted 30.12.19; peer-reviewed by M Hua, T Liyuan; comments to author 15.02.20; revised version received 14.03.20; accepted 19.03.20; published 19.05.20.

Please cite as:
Wang E, Wang F, Yang Z, Wang L, Zhang Y, Lin H, Wang J
A Graph Convolutional Network–Based Method for Chemical-Protein Interaction Extraction: Algorithm Development
JMIR Med Inform 2020;8(5):e17643
URL: http://medinform.jmir.org/2020/5/e17643/
doi:10.2196/17643
PMID:32348257

Original Paper

A Method to Learn Embedding of a Probabilistic Medical Knowledge Graph: Algorithm Development

Linfeng Li[1,2*], PhD; Peng Wang[3,4*], PhD; Yao Wang[2], MSc; Shenghui Wang[1], PhD; Jun Yan[2], PhD; Jinpeng Jiang[2], MSc; Buzhou Tang[5], PhD; Chengliang Wang[3], PhD; Yuting Liu[6], PhD

[1]Institute of Information Science, Beijing Jiaotong University, Beijing, China

[2]Yidu Cloud Technology Inc, Beijing, China

[3]College of Computer Science, Chongqing University, Chongqing, China

[4]Southwest Hospital, Chongqing, China

[5]Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China

[6]School of Science, Beijing Jiaotong University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Yuting Liu, PhD
School of Science
Beijing Jiaotong University
No 3 Shangyuancun Haidian District
Beijing
China
Phone: 86 13810004230
Email: ytliu@bjtu.edu.cn

Abstract

Background: Knowledge graph embedding is an effective semantic representation method for entities and relations in knowledge graphs. Several translation-based algorithms, including TransE, TransH, TransR, TransD, and TranSparse, have been proposed to learn effective embedding vectors from typical knowledge graphs in which the relations between head and tail entities are deterministic. However, in medical knowledge graphs, the relations between head and tail entities are inherently probabilistic. This difference introduces a challenge in embedding medical knowledge graphs.

Objective: We aimed to address the challenge of how to learn the probability values of triplets into representation vectors by making enhancements to existing TransX (where X is E, H, R, D, or Sparse) algorithms, including the following: (1) constructing a mapping function between the score value and the probability, and (2) introducing probability-based loss of triplets into the original margin-based loss function.

Methods: We performed the proposed PrTransX algorithm on a medical knowledge graph that we built from large-scale real-world electronic medical records data. We evaluated the embeddings using link prediction task.

Results: Compared with the corresponding TransX algorithms, the proposed PrTransX performed better than the TransX model in all evaluation indicators, achieving a higher proportion of corrected entities ranked in the top 10 and normalized discounted cumulative gain of the top 10 predicted tail entities, and lower mean rank.

Conclusions: The proposed PrTransX successfully incorporated the uncertainty of the knowledge triplets into the embedding vectors.

(JMIR Med Inform 2020;8(5):e17645) doi:10.2196/17645

KEYWORDS

probabilistic medical knowledge graph; representation learning; graph embedding; PrTransX; decision support systems, clinical; knowledge graph; medical informatics; electronic health records; natural language processing

Introduction

Background

In medical fields, knowledge graphs (KGs) are the core underlying component of a clinical decision support system [1]. Clinical decision support system applications based on KGs have been reported in different scenarios, such as medicine recommendations [2] and drug-to-drug similarity measurements [3]. The KG is a graph-based knowledge representation method, which uses a set of (head, relation, tail) triplets to represent the various entities and their relationships in a domain. Each triplet is called as a fact as well. In KGs, nodes represent entities and edges represent relationships between entities. Medical KGs can be built either by human experts or by using unsupervised data mining from electronic medical records (EMRs). The first approach is too labor-intensive to be feasible for building large-scale KGs. Thus, unsupervised or semisupervised data mining from EMR data is a promising approach [4].

To learn effective knowledge representations, KG embedding has been proposed and gained massive attention, since the embedding vectors are easier to manipulate than the original symbolic entities and relations. The embedding algorithm maps the symbolic entities and relations into a continuous low-dimension vector space while preserving their semantic information. Different approaches to embed KGs by using translation-based learning embedding vectors are reported, such as TransE/H/R/D/Sparse [5-9] (noted as TransX hereafter). TransX algorithms learn embedding vectors from the deterministic facts in a KG. The learned embedding vectors help to improve performance of knowledge completion and other common natural language processing tasks [10]. In the medical field, embedding vectors are reported to be capable of improving diagnostic inference tasks [11].

Related Approaches of Knowledge Graph Embedding

Bordes et al [5] implemented a translation-based algorithm (TransE) to model the (*h*,*r*,*t*) triplets in KGs. The score value of a given triplet is defined as the distance between h+r and t. The margin-based ranking criterion is defined as a loss function, and its target is to make the score value of a positive triplet be lower than that of a negative triplet by some margin. TransE features low model complexity while achieving relatively good predictive performance. Although the criticism has been made that TransE might learn similar vector representations for different tail entities in a 1-to-N relationship, the experiment results for TransH [6] and our study proved that such a flaw is not significant when the number of relations is small enough.

TransH [6] introduced relation-dependent hyperplanes to handle reflexive, 1-to-N, N-to-1, and N-to-N relations. The head and tail embedding vectors are mapped to the relation-dependent hyperplane, making it possible to project one entity into different projection vectors in different relations.

TransR [7] further extends the idea of the relation-specific projection by proposing to project an entity-embedding vector into a relation-specific vector space instead of a hyperplane. The introduction of the relation-specific space makes TransR more expressive at modeling differences among the relation and entities. Thus, TransR surpassed TransH in predicting the tail entities in many-to-many relations. CTransR is an extension of TransR that is designed to handle the differences in each relation. By clustering pairs within 1 relation, the implicit subtypes of a given relation are modeled as a cluster-specific relation vector. In the medical graph in this study, there is no such diversity in the relations.

TransD [8] replaces the relation-specific projection matrix with dynamic projection matrices for each entity-relation pair, thereby modeling various types and attributes among the entities. In addition, the dynamic projection matrices are constructed from projection vectors of head or tail entities and relations, requiring much fewer parameters and resulting in more efficient training. The proportion of corrected entities ranked in the top 10 (Hits@10) of predicting tails in many-to-many relations improved from 73.8% to 81.2% compared with CTransR.

Ji et al [9] argued that previous studies overfit simple relations and underfit complex relations, since relationships are heterogeneous and unbalanced. To resolve the challenge, they proposed the use of sparse matrices (TranSparse), either for each relation—that is, TranSparse(share)—or for each entity and relation—that is, TranSparse(separate). The sparse degree of each projection matrix is determined by the frequency of each relation or relation-entity pair in the training set. TranSparse(share) could be viewed as the counterpart of TransR, while TranSparse(separate) is the counterpart of TransD.

DIST_MULT [12] and ComplEx [13] do not use a distance-based score function; rather, they use a bilinear scoring function within a real vector space or a complex vector space. The loss function of DIST_MULT is the same as that of TransX, mentioned above, while the loss function of ComplEx is the negative log-likelihood of the logistic model. Experimental results show that the bilinear scoring function is more expressive, since DIST_MULT outperforms all TransXs at predicting tail entities in 1-to-N relations. ComplEx further surpassed DIST_MULT because the asymmetry in the products of complex embeddings helps to better express asymmetric relationships. However, because the score value is not distance based, it is difficult to map score values to probabilities.

He et al [14] proposed a density-based embedding method (KG2E) that embeds each entity and relation as a multidimensional Gaussian distribution. Such an embedding method is aimed at modeling the uncertainty of each entity and relation. However, KG2E did not achieve a better Hits@10 when predicting tails in many-to-many relations than CTransR [14]. We note that the uncertainty in KG2E is different from the probabilities of triplets use in this study. In KG2E, an entity is considered to have high certainty if it is contained in more triplets. In contrast, in this study we considered the probability as a metric of certainty of a triplet.

Fan et al [15] proposed a probabilistic belief embedding (PBE) model to measure the probability of each belief (*h*,*r*,*t*,*m*) in large-scale repositories. The notation *m* is the mention of the relation. The problem that PBE tried to solve is the most similar one to the problem we address in our study among all the related studies, which is embedding probability information from KGs into vectors. Apart from the extra element relation mention in

the quad, the key differences in the algorithms are as follows. First, in PBE, the probability of each triplet is calculated not only by using the embedding vectors of its head, relation, and tail, but also by using the embedding vectors of other triplets. In our study, the probability of each triplet depended only on the embedding vectors of itself. Second, in PBE, the softmax function is used to map distances into probabilities. The limitation of using the softmax is that it is difficult to model 1-to-many, many-to-1, and many-to-many relations. Consider that there are multiple tail entities, which are valid tail entities for a given head entity and relation. The ground truth probabilities of these triplets are 1. In Fan and colleagues' equations (8) and (14), the probability values of $Pr(r|h,r)$ are impossible to be trained to 1 for multiple valid tail entities, since the softmax function requires the sum of all probabilities to be 1.

Xiao et al [16] proposed a generative model (TransG) to learn multiple relation semantics. In medical fields, the relations do not contain different meanings as in a general knowledge base; thus, we did not necessarily consider TransG in this study.

Qian et al [17] argued that TransH/R/D/Sparse failed to learn that "various relations focus on different attributes of entities." Their model, TransAt, split the whole entity set into 3 according to the k-means: the capable candidate set of the head, the capable candidate set of the tail, and the rest. Furthermore, it set constraints on the distances of entities in intraset and interset pairs. In the medical KG that we built in our study, the candidate entities of the head and tail for a given relation are clearly defined by the relation itself. Thus, we did not consider TransAt in this study.

In contrast with the deterministic facts in general domain KGs, most facts in the medical domain are probabilistic. For example, considering the triplet (pneumonia, *disease_to_symptom*, fever), the symptom *fever* is common but not always present among patients with the disease *pneumonia* (code J18.901 in the *International Classification of Diseases, Tenth Revision* [ICD-10]). By counting the number of cooccurrences of *fever* and *pneumonia*, the conditional probability $P(symptom = fever|disease = pneumonia)$ could be calculated from a real-world EMR data set.

Such a probabilistic nature is the unique challenge in embedding real-world EMR data-based medical KGs. In the KG on which TransX is designed, the label of each triplet is regarded as absolutely correct or wrong. During the training of TransX, a negative triplet (a triplet with an incorrect label) corresponding to each correct triplet is randomly sampled. Since the labels of triplets are binary in TransX, the training target is set as the score value of the positive triplet being lower than the score value of the negative triplet by some margin. In a medical KG, the label of a triplet—that is, the probability of the triplet—could indicate that (1) one triplet could be more likely (or unlikely) to appear than another or (2) the degree of certainty of each triplet is precisely expressed by the conditional probability. The probability of triplets is not considered by the TransX algorithms.

Objective

We proposed PrTransX, based on the classical TransX algorithms, to learn embedding of a probabilistic KG. The main contributions of this study are as follows.

Mapping Function Between Score Value and Probability

The score values of triplets in the existing TransX algorithms can reflect the geometric distance between the head and tail entities under a given relation. We proposed a function that can map the score value to the probability of a triplet and vice versa. This function both helps in translating the probability of a triplet to a target score during model training phase and enables users to convert a score value to a probability when predicating unforeseen links in model applications.

Loss Function to Learn Embedding of a Probabilistic Knowledge Graph

Based on the mapping function, we introduced probability-based loss of triplets into the original margin-based loss function. Thus, the new loss function requires that the predicted probability of a triplet approximate its statistically probability on a training set. Finally, the triplet probabilistic information is learned into the embedding vectors.

Evaluation of the Proposed Algorithms on Large-Scale Real-World Electronic Medical Records Data

We built probabilistic medical KGs from more than 10 million real-world EMR documents. KG embeddings were learned by the proposed algorithms.

Methods

Notation Overview

Table 1 explains the meanings of all mathematical symbols in this section.

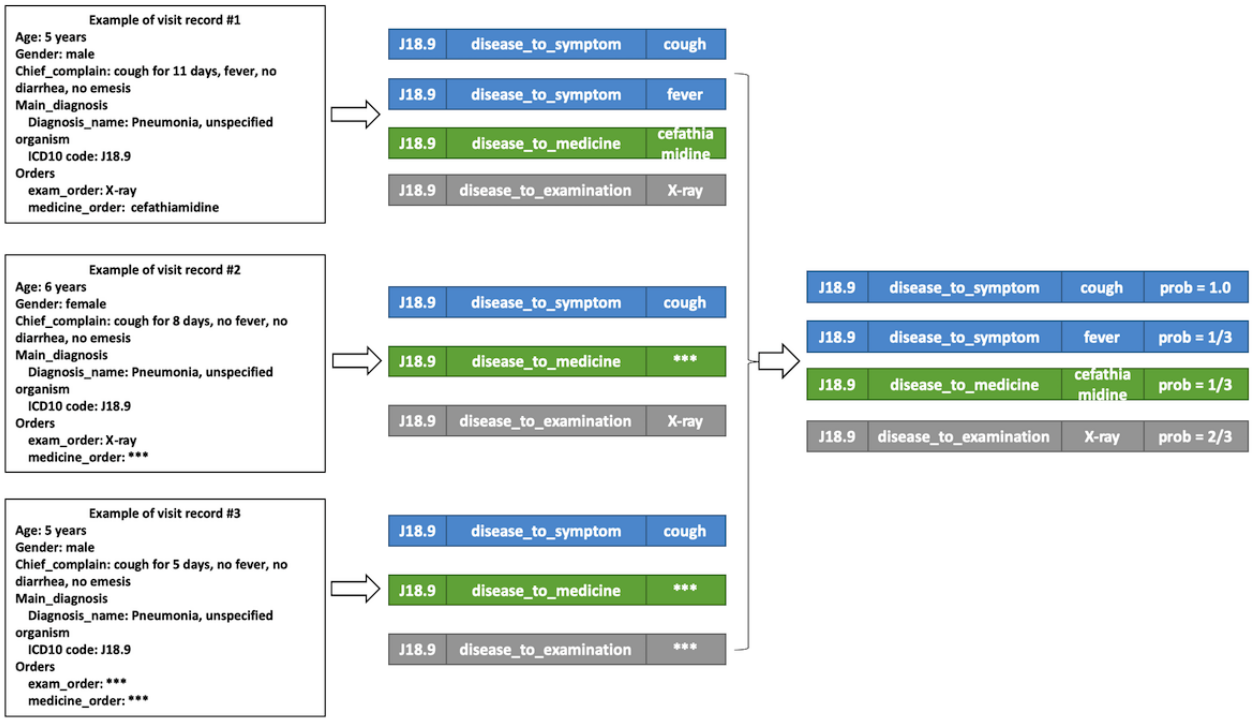
Building the Knowledge Graph From Real-World Electronic Medical Records

Real-world EMR data can be viewed as collections of visit records, each of which consists of all the medical records that are generated within 1 particular visit to a doctor by 1 patient, such as patient information, chief complaint, history of present illness, and medical orders. In each visit record, there are probably multiple medical entities. The term medical entity refers to a concrete instance of diagnosis, symptom, laboratory test, examination, medicine, and operation, such as the diagnosis *pneumonia, unspecified organism*, the symptom *cough*, the medicine *cefathiamidine*, and the examination *x-ray* in Figure 1.

Table 1. Notations used in the study.

Symbols	Meaning
h, r, t, h', r, t'	Head entities, relation, tail entities from positive triplet and negative triplet corresponding to positive triplet (marked as ′)
Δ, Δ'	Set of positive/negative triplets
h, r, t	Embedding vectors of head, relation, and tail entities
h_p, t_p	Projection vectors of head and tail entities
$s_{(h,r,t)}$	Score value of given triplet
$p_{(h,r,t)}$	Probability of given triplet
$\Phi()$, Φ^{-1}	Mapping function between the score value and probability of triplets
$PL_{(h,r,t)}$	Probability-based loss of given triplet
ε_n	Probability value of negative triplet
ε_p	Minimum probability value of positive triplet
λ, K, γ	Scaling factors, margin parameters for loss function
α_r, β_r	Parameters for given relation r
$[x]_+$	The positive part of x
L_m	Margin-based loss function
L	Loss function

Figure 1. Workflow for extracting probabilistic knowledge triplets from real-world electronic medical record data. ICD10: *International Classification of Diseases, Tenth Revision*.



The relationships, which are expressed in (head entity, relation, tail entity) triplets, can be derived from the medical entities that occur in each single visit. By using the example visit above, several triplets can be derived, such as (J18.9, *disease_to_symptom*, cough) or (J18.9, *disease_to_medicine*, cefathiamidine).

The statistical probability of each triplet (h,r,t) can be calculated by the equation $p_{(h,r,t)} = [N_{(h,r,t)}]/N_h$, where $N_{(h,r,t)}$ represents the number of visit records that would derive (h,r,t) triplets, and N_h represents the number of visit records that hold entity h. When (h,r,t) is a valid triplet, the condition $p_{(h,r,t)}$ belongs to the set (0,1] holds; otherwise, $p_{(h,r,t)}$=0.

Data Set Split and Ground Truth Setup

We used the *triplet group* as the minimum unit of separation between the training set and test set. One triplet group contained all triplets that shared the same head and relation. For example, all triplets whose head entity is disease C16.902 (gastric cancer not otherwise specified) and relation is *disease_to_symptom* constitute a triple group, containing (C16.902, *disease_to_symptom*, s_1), (C16.902, *disease_to_symptom*, s_2), ..., and (C16.902, *disease_to_symptom*, s_N), where s_i indicates tail symptoms. One triplet group was split into either the training set or test set. In other words, for any given *h* and *r*, either all its tails were in training data or all were in test data.

In addition to the triplet group, we applied 2 rules during separation. First, for the relation *disease_to_XXX*, we randomly selected 20% of the triplet groups as the test set. Second, for the relation *upper_disease_to_lower_disease*, we included all triplets in the training set, since these relations are prior knowledge.

Noise is inevitable in real-world data. In medical cases, the most common noise in extracted triplets is due to unspecific triplets. Considering E11.901 (type 2 diabetes) as the head entity and *disease_to_laboratory* as the relation, triplets with routine laboratory test items (such as routine blood test) usually have higher probabilities than specific triplets (such as hemoglobin A_{1c}). However, according to medical knowledge, hemoglobin A_{1c} is directly related to E11.901 (type 2 diabetes) as a diagnostic criterion, whereas routine blood test is not. Therefore, the original tail entities in a raw test set may be not really related to a head entity in medicine, and thus evaluating their performance cannot reveal the real ability to predict unknown medical knowledge.

To address the issue, medical experts manually label the ground truth tail entities according to clinical guidelines and the medical literature. During labeling, we also labeled the relevance level for the related tail entities, from 1 of the following values: strongly related (***), related (**), or weakly related (*). Since this is quite labor-intensive, we manually labeled only a randomly selected subset of the test data set. We named this subset the *evaluation data set*.

Negative Triplets Sampling

In TransE [5], the negative triplets are created by randomly replacing the head or tail with a randomly selected entity. The authors of TransH [6] pointed out the false-negative issue and proposed that a Bernoulli distribution should be used to determine whether the head or tail should be replaced, which resulted in a lower number of false-negatives. We used sampling by using a Bernoulli distribution in the following TransH/R/D/Sparse algorithms.

In a medical KG that is built from real-world EMR data, background knowledge provides a more reliable way to sample negative triplets than the previous statistical method.

First, since the types of head and tail entities can be unambiguously determined by their relation, the valid candidate set for the replacement of the head or tail entity can be determined. For example, given the *disease_to_symptom* relation, the head entity must be the diagnosis type and the tail entity must be the symptom type.

Second, we can reduce the false-negatives by using medical knowledge. For example, the disease pneumonia (ICD-10: J18.901) is a hyponym of disease pneumonia unspecified (ICD-10: J18) in the ICD-10 [18] system. If the positive triplet is (pneumonia (ICD-10: J18.901), *disease_to_symptom*, symptom_x), we should not replace the head with any hyponym of pneumonia unspecified (ICD-10: J18) because diseases in the same ICD-10 class are likely to have similar properties.

In summary, by considering the entity type and domain knowledge, the number of false-negative triplets can be effectively reduced, against using the pure statistical method observed in previous work. We tested all the baseline models and proposed models using this enhanced negative sampling approach in our study.

Mapping Function Between the Score Value and Probability

In previous TransX algorithms, the score value of a triplet was defined as the distance between h_p+r and t_p, where h_p and t_p are vectors that are projected from the original embedding vectors h and t by some means, respectively. Such a score value calculation can be viewed as a function of (h,r,t), which is defined as $s_{(h,r,t)} = f_r(h,t) = \|h_p+r-t_p\|_{L1/L2}$.

Based on the above definition, the value range of the score value is [0, +inf). Furthermore, a value of 0 indicates a perfect match among (h,r,t), whereas a positive infinity value indicates that there is no relation between h and t under relation r. In the training of TransX, the target of the loss function was to simultaneously make the score value of the correct triplet of embeddings, which we noted as (h,r,t), as near to 0 as possible and to make the score value of the negative triplet, which we noted as (h′r,t′), as large as possible.

In a medical KG, the label of a triplet is not absolutely correct or wrong, but rather it is a probability value that can be interpreted as the likelihood of the triplet being established. The probability value is 1 if a triplet holds under all conditions, and the probability value is 0 if the triplet does not hold under any valid condition.

It is clear that both the score value and probability can tell how likely a triplet is to be established. They are just different metrics. Therefore, we define a mapping function, Φ(), and its inverse function by Equation (1) in Figure 2.

Figure 2. Equations.

Equation (1)

$$p_{(h,r,t)} = \Phi(f_r(\mathrm{h},\mathrm{t})) = e^{-\lambda \cdot f_r(\mathrm{h},\mathrm{t})}$$
$$f_r(\mathrm{h},\mathrm{t}) = \Phi^{-1}(p_{(h,r,t)}) = \frac{1}{\lambda}\ln\frac{1}{p_{(h,r,t)}}$$

where $\lambda > 0$ is a scaling coefficient

Equation (2)

$$p_{(h,r,t)} = \begin{cases} p_{(h,r,t)} \\ \varepsilon_p \end{cases} \quad \text{if } p_{(h,r,t)} < \varepsilon_p$$

where $0 < \varepsilon_n < \varepsilon_p$

Equation (3)

$$\mathcal{L}_m = \sum_{(h,r,t)\in\Delta}\sum_{(h',r,t')\in\Delta'} [s_{(h,r,t)} + \gamma - s_{(h',r,t')}]_+$$
$$= \sum_{(h,r,t)\in\Delta}\sum_{(h',r,t')\in\Delta'} [f_r(\mathrm{h},\mathrm{t}) + \gamma - f_r(\mathrm{h}',\mathrm{t}')]_+$$

Equation (4)

$$PL_{(h,r,t)} = |\frac{1}{\lambda}\ln\frac{1}{p_{(h,r,t)}} - f_r(\mathrm{h},\mathrm{t})|$$

Equation (5)

$$PL_{(h',r,t')} = [\frac{1}{\lambda}\ln\frac{1}{\varepsilon_n} - f_r(\mathrm{h}',\mathrm{t}')]_+$$

Equation (6)

$$\mathcal{L} = \sum_{(h,r,t)\in\Delta}\sum_{(h',r,t')\in\Delta'} (e^{\frac{1}{K}[s_{(h,r,t)}+\gamma-s_{(h',r,t')}]_+}) \cdot (\alpha_r \cdot PL_{(h,r,t)} + \beta_r \cdot PL_{(h',r,t')})$$

where the hyperparameter K is a scaling factor to adjust the weight of the margin-based loss value; γ is the margin used to calculate the margin-based loss in the TransX model; α_r and β_r are nonnegative coefficients used to adjust the weights between the positive probability-based loss and negative probability-based loss; the subscript r in α_r and β_r means that these hyperparameters are set to different values for different relations.

To avoid dividing by 0, for a negative triplet t ($p_{(h'r,t')} = 0$), we defined an approximation as $p_{(h'r,t')} \approx \varepsilon_n, \varepsilon_n > 0$. To avoid the scenario that the probability of a valid triplet (h,r,t) is less than ε_n, we introduced the minimum probability of valid triplet as ε_p in Equation (2), Figure 2).

The mathematical properties of $\Phi(d)$ facilitate learning effective embedding vectors from a medical KG. First, we 1-to-1 mapped the probability value in {x| belongs to the set $[\varepsilon_n,1]$} and the distance value in {d|d belongs to the set $[0,\ln(1/\varepsilon_n]$}. This made it possible to infer the quantitative probability of an unseen triplet only using the embedding vectors. Second, the calculation of the probability of (h,r,t) only depends on the embedding vectors (h,r,t) and not on any other embedding vectors. This means that the probability of one triplet could be independent of another. In 1-to-many and many-to-1 relations, it is common that multiple triplets are valid with high probabilities, and such independence avoids the shortcomings of using the softmax function to calculate the probability of a triplet in PBE [15].

Loss Function to Learn Embedding of the Probabilistic Knowledge Graph

In previous TransX algorithms, the margin-based ranking criterion is defined as a loss function, which is expressed as Equation (3) in Figure 2. The training objective was to minimize the value of the loss function.

The major shortcoming of such a margin-based loss function is that it only requires $f_r(\mathrm{h},\mathrm{t})+\gamma \le f_r(\mathrm{h}'\mathrm{t}')$, but it does not require that the predicted probability of each triplet approximate its statistical probability. To address the shortcoming, we defined the *probability-based loss of triplet*. Given a triplet (h,r,t) with probability $p_{(h,r,t)}$, we required that the targeted score value approximate the mapping of the actual probability $p_{(h,r,t)}$ by $\Phi()$. Thus, we defined the loss as in Equation (4) in Figure 2. We defined the loss of the negative triplet by Equation (5) in Figure 2. We defined the loss function L of PrTransX as a combination of the margin-based loss, the probability-based loss of positive triplets, and the probability-based loss of negative triplets; see Equation (6) in Figure 2.

Evaluation Protocol

To evaluate the performances of different algorithms, we used the link prediction task. The objective of link prediction in this study was to predict the tail entities using the heads and relations on the evaluation data set. For each triplet group in the evaluation data set, the embedding vectors that are trained by each algorithm can predict a sequence of tail entities in a valid entity type. An algorithm is considered to be better than another if it shows better results on evaluation metrics.

Referring to Bordes et al [5], we used Hits@10 and mean rank (the mean of those correctly predicted ranks) as evaluation metrics. The term *correctly predicted* or *hit* refers to the predicted item existing in the ground truth tail entities, regardless of the relevance level. Given that different tail entities would have different relevance levels with the head entities in medical knowledge, we also used the normalized discounted cumulative gain of the top 10 predicted tail entities (NDCG@10) [19] as a metric, since it measures whether an algorithm can rank more relevant tail entities in the front.

Real-World Electronic Medical Records Data and Knowledge Graph

We performed this study using a dataset from the data platform and application platform of the Southwest Hospital in China. The platform is built based on a distributed computing architecture, and it is located on a private cloud in the hospital. The data platform and application platform aggregates medical data from EMRs, the hospital information system, laboratory information system, picture archiving and communication systems, and other isolated subsystems. The platform organizes all related medical data into visit-level and patient-level data. Visit-level data contain records from all subsystems for the same visit, such as the chief complaints and present illness histories from EMRs, examination orders and drug prescriptions from the hospital information system, and laboratory examination results from the laboratory information system. Patient-level data contain all visit data of the same patient.

Results

Data Set

We collected EMR records from 2015 to 2018 in the data set. The data set consisted of 3,767,198 patients and 16,217,270 visits. The entities from the data set were in 6 categories, described in Table 2. Among the entities, the disease entity was identified by its unique ICD-10 code [18]. ICD-10 hierarchical relationships are considered when extracting triplets. For example, assuming that the main diagnosis of a visit is C16.902 (gastric cancer not otherwise specified) and medicine m1 is prescribed by doctor, a list of triplets will be generated: (C16.902, *disease_to_medicine*, m1), (C16.9, *disease_to_medicine*, m1), and (C16, *disease_to_medicine*, m1).

Table 2. Description and distribution of relationships in the medical knowledge graph.

relation_name	Source	Head entity type	Tail entity type	Triplet count
disease_to_medicine	EMR[a] data set	Disease	Medicine	74,835
disease_to_symptom	EMR data set	Disease	Symptom	53,885
disease_to_operation	EMR data set	Disease	Operation	13,292
disease_to_laboratory	EMR data set	Disease	Laboratory	71,805
disease_to_examination	EMR data set	Disease	Examination	38,061
upper_disease_to_lower_disease	Domain knowledge: ICD-10[b]	Disease	Disease	6455
Total	—[c]	—	—	258,333

[a]EMR: electronic medical record.

[b]ICD-10: International Classification of Diseases, Tenth Revision.

[c]Not applicable.

Based on the data set, we extracted triplets of 5 relationships. We added the relation *upper_disease_to_lower_disease* to the relation list as prior domain knowledge. Table 2 describes the features of the relationships. The original number of triplets in the data set was quite large because the probability of triplets was distributed as a long-tailed distribution. To reduce the noise from data and improve the training efficiency by reducing training data, we selected only the top 20 triplets (sorted in descending order of probability).

After separation, there were 205,877 triplets from 21,327 triplet groups in the training set, and 49,756 triplets from 4547 triplet groups in the test set.

There were in total 335 triplets from 25 triplet groups in the evaluation data set. We distributed the 25 triplet groups evenly among the 5 relationships to be evaluated.

Baselines and Implementation

All translational distance-based TransX models could be extended to be PrTransX. Thus, the evaluation target was to compare the performances of the TransX methods and their corresponding PrTransX methods. We selected some of the translational distance-based TransX models: TransE [5], TransH [6], TransR [7], TransD [8], and TranSparse [9].

In addition to TransX, we also examined whether probabilistic inference algorithms, such as the naive Bayes, are feasible in this application. The conclusion was that they were not. To

predict tail entities—that is, to calculate $P(t = t_i|h = h_0, r = r_0)$ for each candidate tail entity—the conditional likelihood $P(h = h_0|t = t_i, r = r_0)$ must be available from the training set. However, since the data split of the triplet group ensures that no triplet in the form of (h_0,r_0,t) exists in the training set, the conditional probability is not available.

In the experiments, we set the size of the embedding vectors of all entities and relations to be 20. We used the L1 distance in training. We set the other parameters as follows for all models: K=1000, γ=1, λ=10, α_r={1,1,1,1,1,10}, β_r={15,10,10,15,10,0}, $\varepsilon_p=10^{-4}$, and $\varepsilon_n=10^{-13}$. The values of each relation of α_r and β_r are in the sequence of *disease_to_medicine*, *disease_to_symptom*, *disease_to_operation*, *disease_to_laboratory*, *disease_to_examination*, and *upper_disease_to_lower_disease*.

Evaluation Results

Figures 3, 4, and 5 compare the performance of TransX algorithms and the corresponding PrTransX algorithms, in terms of Hits@10, mean rank, and NDCG@10. The score for each algorithm is the average value on all of the triplets in the evaluation set. In our study, PrTransX performed better than its corresponding TransX model for X=E/H/R/D/Sparse in nearly all performance indicators, achieving a higher Hits@10 and NDCG@10 and lower mean rank.

Figure 3. Proportion of corrected entities ranked in the top 10 of the tested algorithms.

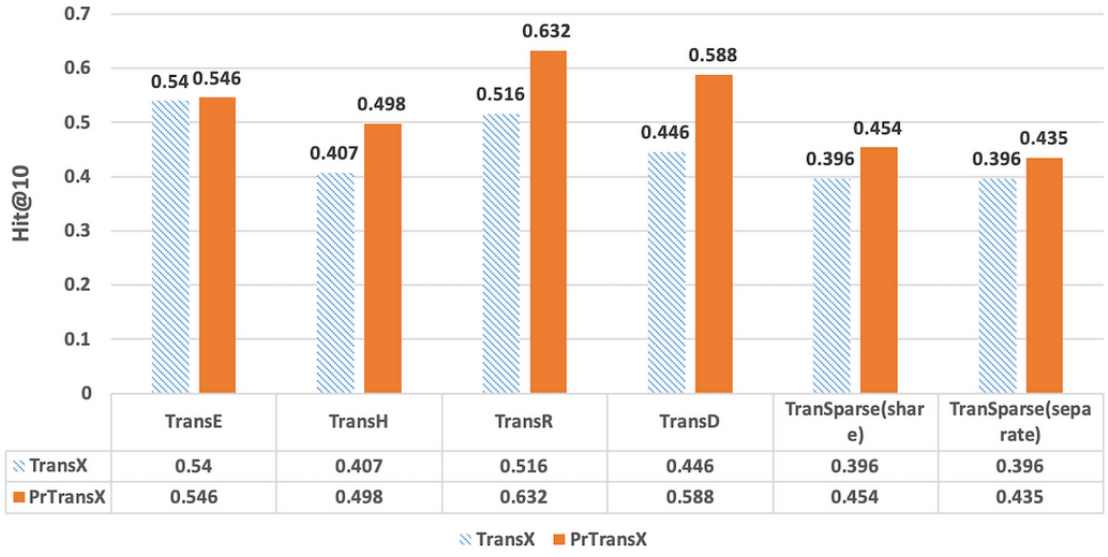


Figure 4. Mean rank of the tested algorithms.

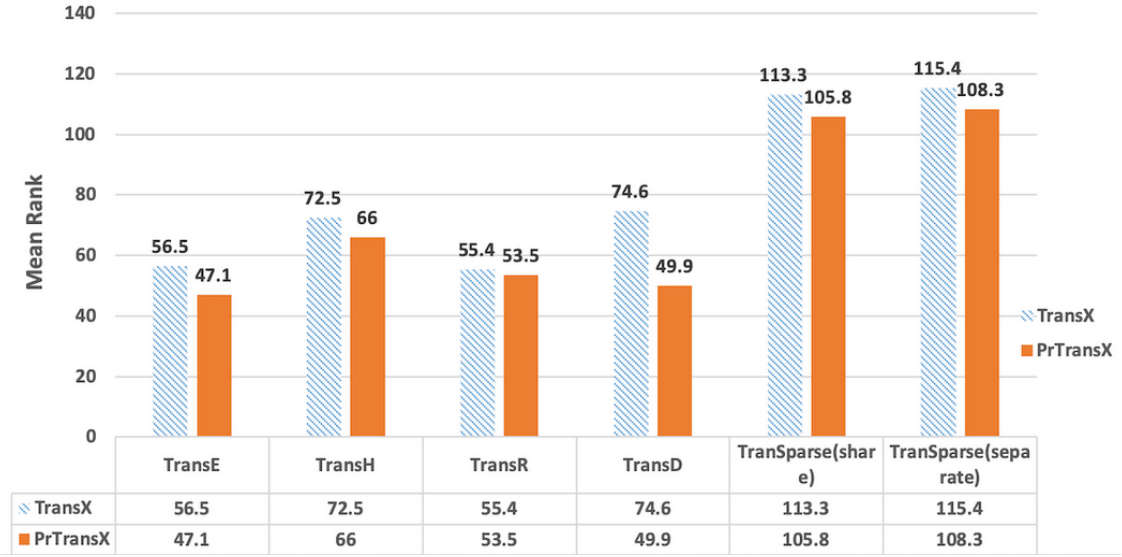
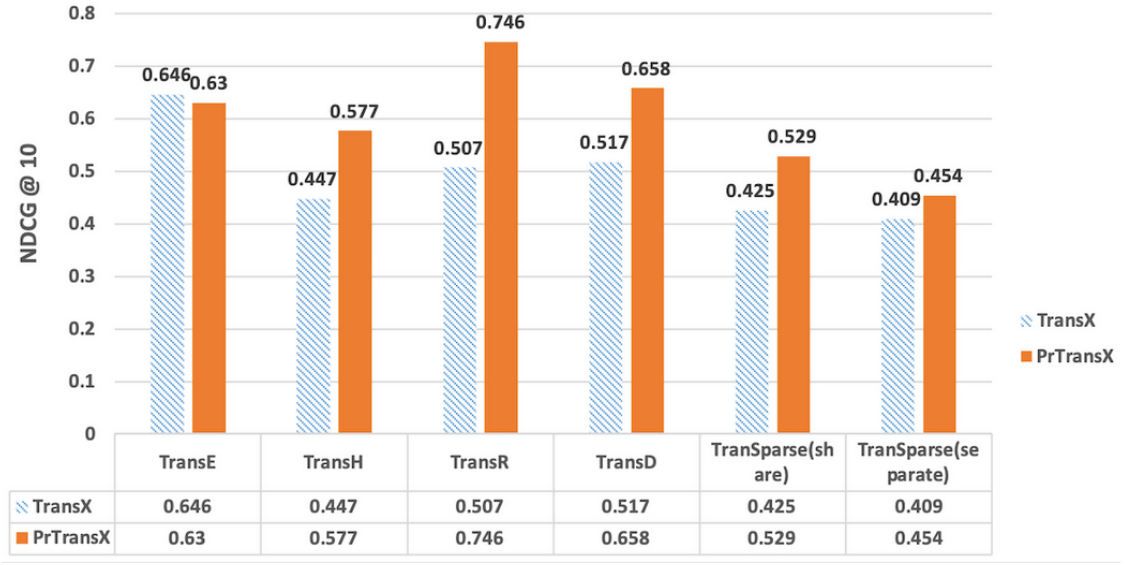


Figure 5. Normalized discounted cumulative gain of the top 10 predicted tail entities of the tested algorithms.



In addition to comparing the average score on the triplets for all the relations, we also examined the detailed performances for each relation.

In order to describe the distribution characteristics of entities by each relation in a quantitative manner, we calculated the *number of head entities per tail entity* by relations. For a given relation and tail entity, the *number of head entities per tail entity* is the number of distinct head entities that relate to the tail entity. Next, the numbers of head entities for each of the tail entities for a given relation form *the distribution of head entities numbers* for the relation.

Figure 6 illustrates the distribution of head entities numbers by different relations. For each box, the upper bound, the median line with a notch, and the lower bound represent the value at the 75%, 50%, and 25% percentiles of the data, respectively. The median number of relation *disease_to_examination* is near 150, and the 75% percentile value is above 350, which means that 50% of examination entities are related to near 150 diseases and 25% are related to more than 350 diseases. The average numbers of related diseases for other entities (ie, laboratory, symptom, medicine, and operation) are all less than the examination entity. The one with the fewest is operation.

Figure 6. Distribution of head entities numbers by different relations.

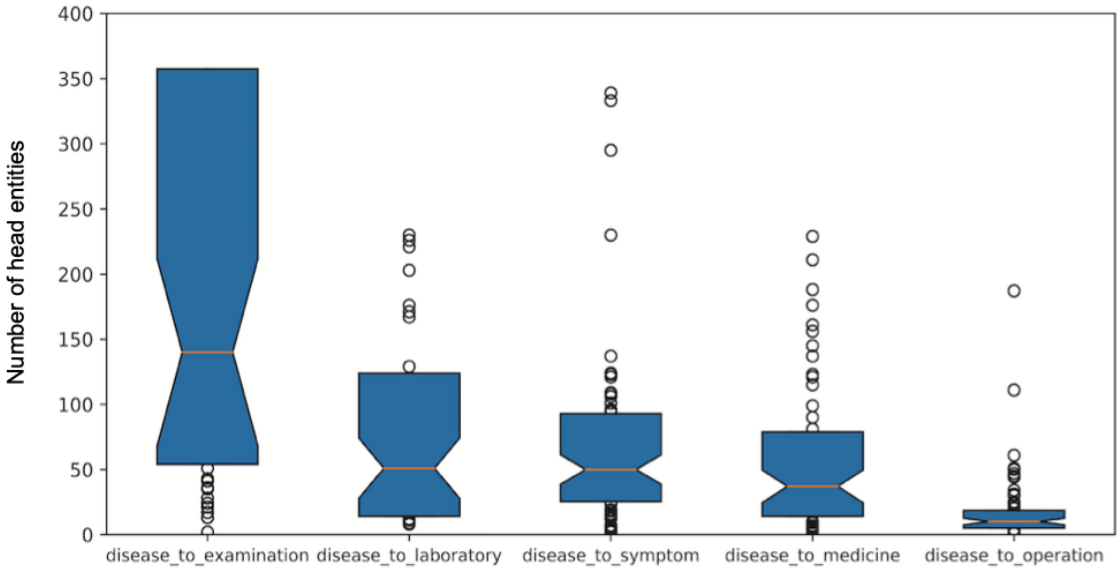
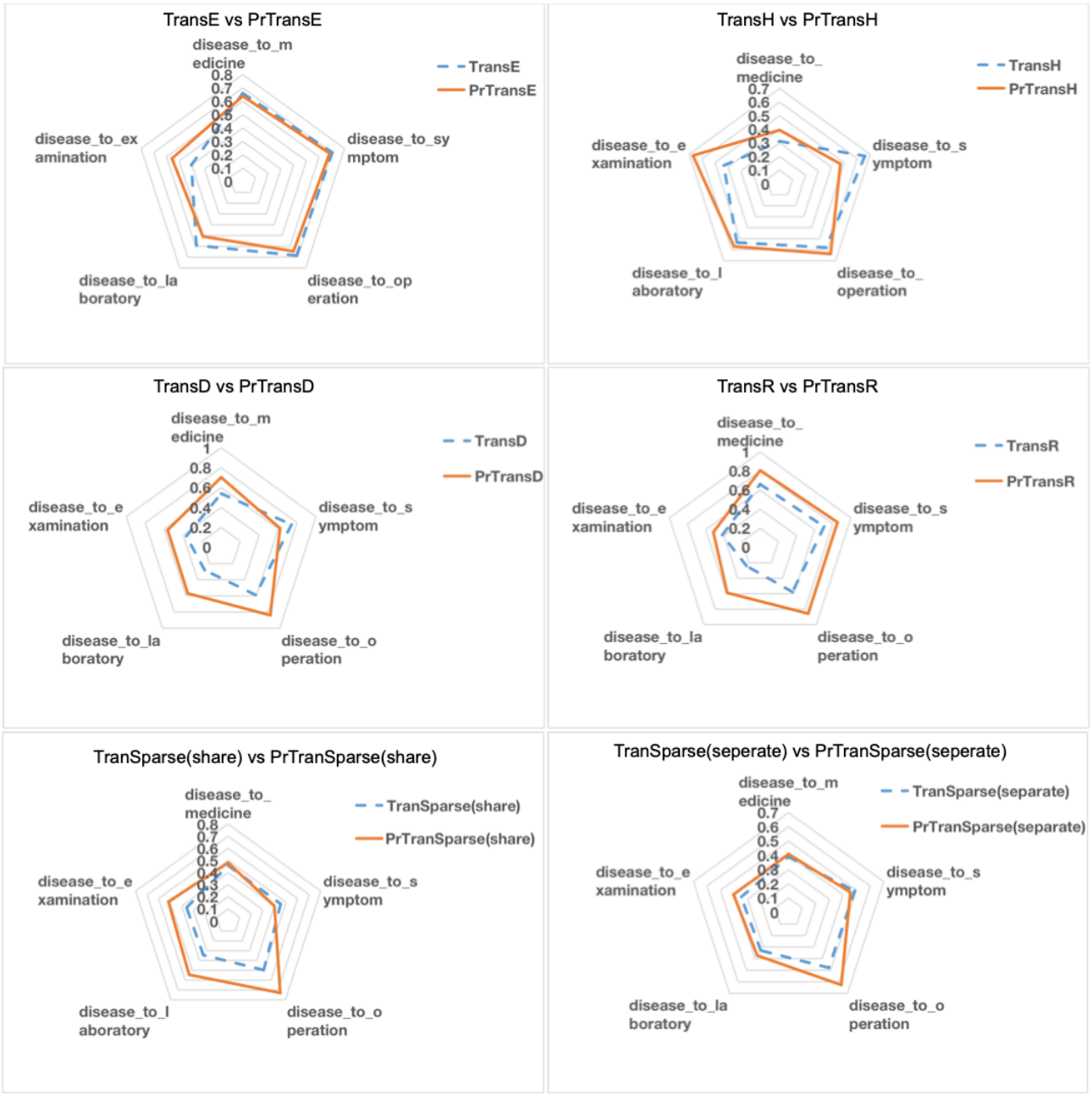


Figure 7 compares NDCG@10 of TransX with its corresponding PrTransX by different relations. Each subgraph represents the NDCG@10 comparison of TransX and PrTransX for a given X model on each of the 5 relations. The performance of PrTransX is drawn as a solid line, while the performance of TransX is drawn as a dashed line.

Figure 7. Comparison of normalized discounted cumulative gain of the top 10 predicted tail entities of TransX and PrTransX by different relations.



Discussion

Principal Findings

As Figures 3, 4, and 5 show, PrTransX performed better than the TransX model for X=E/H/R/D/Sparse in nearly all performance indicators, achieving a higher Hits@10 and NDCG@10 and lower mean rank. All these data proved that by adding probability-based loss into the translation-based loss functions, the obtained embedding vectors improved the link prediction performance. Furthermore, PrTransR performed the best across all PrTransX algorithms under Hits@10 and NDCG@10, and PrTransE performed better under mean rank.

The performance of TransE was generally better than that of TransH/R/D. This was due to the property of our medical KG. In the entire test set, there were only 5 different relations. The number of relations was so small that it was possible to train the embedding of all entities and relations in the same space to satisfy the training objective, which was similar to the result that the link prediction performance of TransH was worse than TransE on the WN18 data set used by Wang et al [6].

The results of TransE and PrTransE were quite similar under the Hits@10 and NDCG@10. In particular, the NDCG@10 of TransE was slightly better than that of PrTransE. Such similar performances were because TransE embedded all entities and relations into the same space. The probability-based training targets of the different relations were impossible to satisfy in the same space. The results of PrTransH/R/D show that the probability-based training targets introduced significant improvements in the Hits@10 and NDCG@10 over those measures of the TransH/R/D model.

These statistics in Figure 6 indicate that the relation *disease_to_examination* was not specific, since each examination entity was related to huge amount of disease entities on average. In contrast, *disease_to_operation* was the most specific relation, as the median number was 10. This conforms the medical common sense that the same examination is applicable to many types of diseases, but the same operation is applicable to only a few types of diseases.

Regarding the obvious performance difference for different relations shown in Figure 7, we can conclude that if the tail entity type is not specific (eg, examination), the link prediction on the evaluation data set is relatively difficult, since it is hard to train tail entities to fit distance requirements of a large amount of distinct head entities. Otherwise, if the tail entity type is specific to disease (eg, operation), the link prediction task would be relatively easier because the training objective is easier to be satisfied.

Overall, the polygons with solid lines (PrTransX) are larger than those of dashed lines (TransX), which means that PrTransX performed better. First, the best NDCG@10 score was achieved by PrTransR in the relation *disease_to_operation*. Moreover, in relation to *disease_to_operation*, PrTransD/Sparse(share)/Sparse(separate) also achieved significant improvements over the respective TransX. Such results echo the previous argument that the link prediction task would be relatively easier for specific relations than for unspecific relations. Second, even if the relation *disease_to_examination* is unspecific (difficult to predict), PrTransX outperformed TransX in this relation for all models.

Limitations

This study had several limitations. First, the algorithm PrTransX can only obtain performance improvement on a probabilistic KG. For the KGs in general domains that have no probability over the relations, PrTransX algorithms are the same as the classical TransX algorithms. Second, the evaluation data set used by the link prediction task contained only 335 triplets from 25 triplet groups, since the labeling is quite labor-intensive.

Future Studies

It should be noted that the application scenarios of knowledge embedding are far beyond link prediction. In state-of-the-art natural language processing research, such as ERNIE [10], significant improvements in various knowledge-driven tasks are achieved by using both text contexts and KGs to train word-embedding vectors. In the medical field, Zhao et al [11] also reported that embedding vectors from TransE and latent factor models could be used in a conditional random field to infer possible diagnoses based on laboratory test results and symptoms. Applying embedding vectors learned by PrTransX from a probabilistic medical KG into the clinical decision support system is worth exploring in the future.

Conclusion

We proposed PrTransX to learn the embedding vectors of a probabilistic KG. We performed the study on a medical KG constructed from large-scale EMR data, and evaluation on link prediction indicated that the embedding learned by the PrTransX significantly outperformed that learned by corresponding TransX algorithms. We can conclude that the proposed PrTransX successfully incorporated the uncertainty of the knowledge triplets into the embedding vectors.

Conflicts of Interest

None declared.

References

1. Kong G, Xu DL, Yang JB. Clinical decision support systems: a review on knowledge representation and inference under uncertainties. Int J Comput Intell Syst 2008;1(2):159-167. [doi: 10.2991/jnmp.2008.1.2.6]
2. Wang M, Liu M, Liu J, Wang S, Long G, Qian B. Safe medicine recommendation via medical knowledge graph embedding. arXiv preprint arXiv:1710.05980. 2017 Oct 26. URL: https://arxiv.org/pdf/1710.05980.pdf [accessed 2020-03-30]
3. Shen Y, Yuan K, Dai J, Tang B, Yang M, Lei K. KGDDS: a system for drug-drug similarity measure in therapeutic substitution based on knowledge graph curation. J Med Syst 2019 Mar 05;43(4):92. [doi: 10.1007/s10916-019-1182-z] [Medline: 30834481]
4. Rotmensch M, Halpern Y, Tlimat A, Horng S, Sontag D. Learning a health knowledge graph from electronic medical records. Sci Rep 2017 Jul 20;7(1):5994 [FREE Full text] [doi: 10.1038/s41598-017-05778-z] [Medline: 28729710]
5. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. San Diego, CA: Neural Information Processing Systems Foundation, Inc; 2013 Presented at: Advances in Neural Information Processing Systems 26; Dec 5, 2013; San Diego, CA p. 2787-2795.
6. Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. 2014 Presented at: Twenty-eighth AAAI Conference on Artificial Intelligence; Jul 27-31, 2014; Quebec, QC, Canada p. 27-31.
7. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. 2015 Presented at: Twenty-ninth AAAI Conference on Artificial Intelligence; Jan 25-30, 2015; Austin, TX, USA.
8. Ji G, He S, Xu L, Liu K, Zhao J. Knowledge graph embedding via dynamic mapping matrix. Stroudsburg, PA: Association for Computational Linguistics; 2015 Presented at: 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers); Jul 26-31, 2015; Beijing, China p. 26-31. [doi: 10.3115/v1/p15-1067]

9. Ji G, Liu K, He S, Zhao J. Knowledge graph completion with adaptive sparse transfer matrix. 2016 Presented at: Thirtieth AAAI Conference on Artificial Intelligence; Feb 12-17, 2016; Phoenix, AZ, USA p. 12-17.
10. Zhang Z, Han X, Liu Z, Jiang X, Sun M, Liu Q. ERNIE: enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129. 2019 Jun 04. URL: https://arxiv.org/pdf/1905.07129.pdf [accessed 2020-03-30]
11. Zhao C, Jiang J, Guan Y, Guo X, He B. EMR-based medical knowledge representation and inference via Markov random fields and distributed representation learning. Artif Intell Med 2018 May;87:49-59. [doi: 10.1016/j.artmed.2018.03.005] [Medline: 29691122]
12. Yang B, Yih W, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. arXiv preprint arXiv:1412.6575. 2015 Aug 29. URL: https://arxiv.org/pdf/1412.6575.pdf [accessed 2020-03-30]
13. Trouillon T, Welbl J, Riedel S, Gaussier É, Bouchard G. Complex embeddings for simple link prediction. 2016 Presented at: 33rd International conference on machine learning; Jun 19-24, 2016; New York, NY, USA p. 19-24.
14. He S, Liu K, Ji G, Zhao J. Learning to represent knowledge graphs with Gaussian embedding. 2015 Presented at: ACM International Conference on Information and Knowledge Management; Oct 19-23, 2015; Melbourne, Australia p. 19-23. [doi: 10.1145/2806416.2806502]
15. Fan F, Zhou Q, Abel A, Zheng TF, Grishman R. Probabilistic belief embedding for knowledge base completion. arXiv preprint arXiv:1505.02433. 2015 May 22. URL: https://arxiv.org/pdf/1505.02433.pdf [accessed 2020-03-30]
16. Xiao H, Huang M, Zhu X. TransG: a generative model for knowledge graph embedding. Stroudsburg, PA: Association for Computational Linguistics; 2016 Presented at: 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Aug 7-12, 2016; Berlin, Germany p. 2316-2325. [doi: 10.18653/v1/p16-1219]
17. Qian W, Fu C, Zhu Y, Cai D, He X. Translating embeddings for knowledge graph completion with relation attention mechanism. 2018 Presented at: 27th International Joint Conferences on Artificial Intelligence; Jul 13-19, 2018; Stockholm, Sweden p. 4286-4292. [doi: 10.24963/ijcai.2018/596]
18. World Health Organization. International Statistical Classification of Diseases and Related Health Problems. Volume 1. Geneva, Switzerland: WHO; 2004.
19. Järvelin K, Kekäläinen J. Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 2002 Oct;20(4):422-446. [doi: 10.1145/582415.582418]

Abbreviations

EMR: electronic medical record
HITS@10: proportion of corrected entities ranked in the top 10
ICD-10: International Classification of Diseases, Tenth Revision
KG: knowledge graph
NDCG@10: normalized discounted cumulative gain of the top 10 predicted tail entities
PBE: probabilistic belief embedding

Edited by T Hao, B Tang, Z Huang; submitted 31.12.19; peer-reviewed by B Qian, C Zhou; comments to author 02.03.20; revised version received 17.03.20; accepted 19.03.20; published 21.05.20.

Please cite as:
Li L, Wang P, Wang Y, Wang S, Yan J, Jiang J, Tang B, Wang C, Liu Y
A Method to Learn Embedding of a Probabilistic Medical Knowledge Graph: Algorithm Development
JMIR Med Inform 2020;8(5):e17645
URL: https://medinform.jmir.org/2020/5/e17645
doi:10.2196/17645
PMID:32436854

Original Paper

Document-Level Biomedical Relation Extraction Leveraging Pretrained Self-Attention Structure and Entity Replacement: Algorithm and Pretreatment Method Validation Study

Xiaofeng Liu[1], PhD; Jianye Fan[1], MD; Shoubin Dong[1], PhD

Communication and Computer Network Key Laboratory of Guangdong, School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

Corresponding Author:
Shoubin Dong, PhD
Communication and Computer Network Key Laboratory of Guangdong
School of Computer Science and Engineering
South China University of Technology
No. 381, Wushan Road
Tianhe District, Guangdong
Guangzhou, 510610
China
Phone: 86 15625125397
Email: sbdong@scut.edu.cn

Abstract

Background: The most current methods applied for intrasentence relation extraction in the biomedical literature are inadequate for document-level relation extraction, in which the relationship may cross sentence boundaries. Hence, some approaches have been proposed to extract relations by splitting the document-level datasets through heuristic rules and learning methods. However, these approaches may introduce additional noise and do not really solve the problem of intersentence relation extraction. It is challenging to avoid noise and extract cross-sentence relations.

Objective: This study aimed to avoid errors by dividing the document-level dataset, verify that a self-attention structure can extract biomedical relations in a document with long-distance dependencies and complex semantics, and discuss the relative benefits of different entity pretreatment methods for biomedical relation extraction.

Methods: This paper proposes a new data preprocessing method and attempts to apply a pretrained self-attention structure for document biomedical relation extraction with an entity replacement method to capture very long-distance dependencies and complex semantics.

Results: Compared with state-of-the-art approaches, our method greatly improved the precision. The results show that our approach increases the F1 value, compared with state-of-the-art methods. Through experiments of biomedical entity pretreatments, we found that a model using an entity replacement method can improve performance.

Conclusions: When considering all target entity pairs as a whole in the document-level dataset, a pretrained self-attention structure is suitable to capture very long-distance dependencies and learn the textual context and complicated semantics. A replacement method for biomedical entities is conducive to biomedical relation extraction, especially to document-level relation extraction.

(JMIR Med Inform 2020;8(5):e17644) doi:10.2196/17644

KEYWORDS

self-attention; document-level; relation extraction; biomedical entity pretreatment

Introduction

A large number of biomedical entity relations exist in the biomedical literature. It is beneficial for the development of biomedical fields to automatically and accurately extract these relations and form structured knowledge. Some biomedical datasets have been proposed for extracting biomedical relations, such as drug-drug interactions (DDI) [1], chemical-protein relations (CPR) [2], and chemical-induced diseases (CID) [3]. The former 2 datasets are sentence-level annotated datasets that

extract relations on a single sentence containing a single entity-pair mention, and the latter is a document-level annotated dataset, which means that it is uncertain whether relations are asserted from within sentences or across sentence boundaries.

Most approaches [4-7] have focused on single sentences containing biomedical relations. For example, Zhang et al [4] presented a hierarchical recurrent neural network (RNN) to combine raw sentences with their short dependency paths for a DDI task. To deal with long and complicated sentences, Sun et al [5] separated sequences into short context subsequences and proposed a hierarchical recurrent convolutional neural network (CNN). Because these approaches cannot be directly applied to document-level datasets, some existing methods [8,9] divided the document-level dataset into 2 parts and trained an intrasentence model and an intersentence model. Nevertheless, because of long-distance dependencies and co-references, their methods cannot be adapted to cross-sentence relation extraction. Furthermore, splitting the dataset resulted in noise and rule-based mistakes.

Currently, for intersentence relation extraction, some studies [10-12] generate dependency syntax trees within sentences and across sentences and employ a graph neural network to capture dependencies. However, it is costly to build dependency syntax trees. In addition, few studies, except those by Li et al [13] and Verga et al [14], have considered the influence of noisy data due to the segmentation of datasets and taking advantage of the textual context. For a document-level annotated corpus, an entity-pair mention within sentences or across sentences has a biomedical relationship by thinking simply, which will undoubtedly cause errors and may ignore plenty of useful information such that many sentences with co-occurring or co-referential medical entity mentions refer to biomedical relations.

For example, the chemical-disease relation (CDR) dataset is a document-level corpus designed to extract CID relations from biomedical literature [15]. For CID relation extraction, most current methods [8,16,17] divide the CDR dataset into intrasentence-level and intersentence-level relation instances using heuristic rules. Although these heuristic rules are effective, they inevitably generate noisy instances of CID relations or ignore some useful information. For example, the following sentence expresses CID relations between the chemical amitriptyline and the disease blurred vision: “The overall incidence of side effects and the frequency and severity of blurred vision, dry mouth, and drowsiness were significantly less with dothiepin than with amitriptyline.”

According to heuristic rules [8], the token distance between two mentions in an intrasentence level instance should be <10. The token distance between the chemical amitriptyline and the disease blurred vision in this example is 12; therefore, this sentence is discarded. However, factually, this sentence is the only sentence in the document [18] that describes the CID relation between the chemical amitriptyline and the disease blurred vision. Obviously, heuristic rules cannot precisely partition the CDR dataset, and they can induce the wrong classification by models, although they use multi-instance learning to reduce these errors.

Therefore, when constructing relation instances from a document-level dataset, it is necessary to consider sentences with multiple mentions of target entities in the entire document. While treating all target entities in a document as a whole brings benefits, the challenges are very long-distance dependencies and complex semantics, from which traditional neural networks such as CNN or RNN cannot accurately extract document-level relations. Recently, pretrained self-attention structures, such as SciBERT [19] and BERT [20], were proposed and were not necessarily better than RNN at capturing long-range dependencies. However, they performed better at increasing the number of attention heads [21]. A pretrained transformer has already learned more semantic features, and it performs well for sentence-level relation extraction; however, it did not apply to document-level relation extraction.

To address these problems, this paper proposes a pretrained self-attention mechanism and entity replacement method to extract document-level relationships. In this way, this paper has several contributions. First, to avoid errors by dividing the document-level dataset, this paper proposes a new data preprocessing method that treats the target entity pair of some sentences in a document as an instance. Second, to better focus on the target entity pairs and their context, a replacement method is proposed to replace biomedical entity pairs with uniform words. Compared with the different entity preprocessing for biomedical entity pairs, the replacement method is more effective for biomedical relation extraction. Third, to solve the problem of long-distance dependencies and learning complex semantics, a pretrained self-attention structure is proposed for document-level relation extraction and to achieve superior performance than state-of-the-art approaches. Through analysis and visualization of the model structure, the effectiveness of the self-attention structure for document-level datasets is demonstrated.

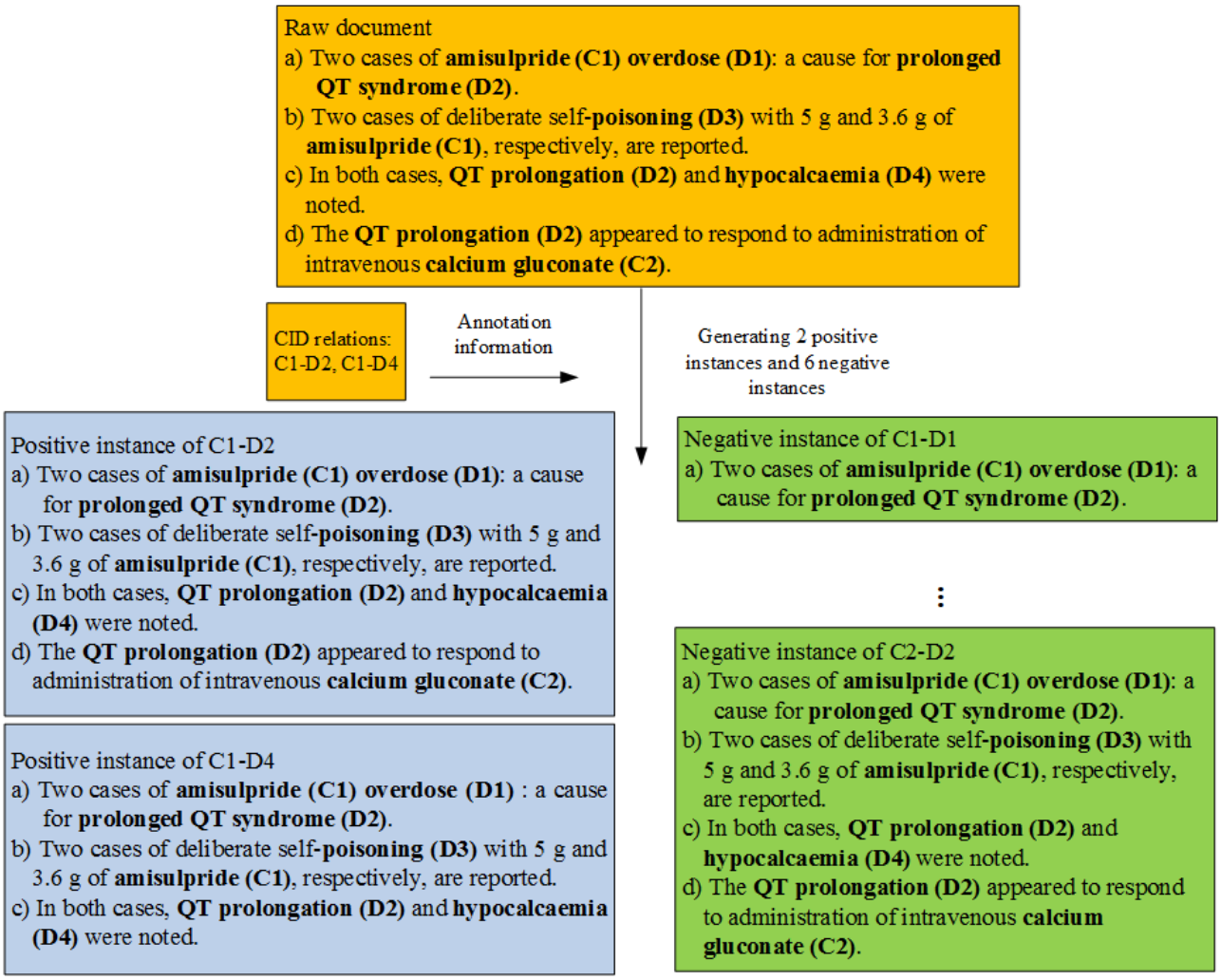
Methods

Data Preprocessing for the Document-Level Corpus

As already mentioned, splitting the document-level corpus will increase noise and may lose some useful information. To address this problem, the sentences in which the target entity pair is located and the sentences between them are constructed to an instance. This approach has the following benefits. First, it does not introduce error messages. The sentences do not need to be labeled after the segmentation of the dataset. The relationship between the marked relation pairs in the document corresponds to the instances one by one. Second, it discards useless information that is not related to the relationship of the target entities. Some are not related to those sentences in which the target entities are located; hence, they are noise for relation extraction. Discarding them will focus the model on the sentences in which the entity pair is located. Third, it keeps a lot of useful information, such as contextual information about entities and the relationship of entities.

As shown in Figure 1, a document [22] in the CDR dataset is constructed into biomedical relation instances. All chemical or disease entities are bold.

Figure 1. An example of document-level relation instance construction.



In this document, there are 2 chemical entities, “amisulpride” (C1) and “calcium gluconate” (C2), and 4 disease entities: “overdose” (D1), “prolonged QT syndrome/QT prolongation” (D2), “poisoning” (D3), and “hypocalcemia” (D4). It should be noted that C1, C2, D1, D2, D3, and D4 are added to the document to indicate which are chemical entities and which are disease entities. Hence, the document can be constructed into 8 instances, in which 2 instances of C1 and D2 or C1 and D4 have CID relations. a), b), c), and d) conformed the instance of C1 and D2. a), b), and c) conformed the instance of C1 and D4.

Semantically, both the intralevel sentence a) and the interlevel sentences b) and c) express the CID relationship of C1 and D2. However, according to heuristic rules, b) and c) will be discarded because only the entities that are not involved in any intrasentence level instance are considered at the intersentence level. Third, instances are full of contextual information of chemical and disease entities, which is conducive to document-level relation extraction when exploiting it well.

There are lots of biomedical entities in a document. When constructing the instances of the target entity pair, it is inevitable that the same instance is tagged with different labels, resulting in incorrect classification. For example, as mentioned in the Methods section, the instances of C1-D2 and C2-D2 are the same but tagged with different labels. To solve this problem, entity pretreatment methods are presented.

There are 2 different biomedical entity pretreatments, as shown in Figure 2. In the first pretreatment, the target chemical and disease entities are respectively replaced with “chemical” and “disease,” which are called the replacement method. For example, in the instance of C1 and D2, sentence a) will be processed into “Two cases of chemical entity: a cause for disease.” In addition to the replacement method, there is another data preprocessing method, called the addition method. Different marks are added to the boundaries of chemical and disease entities, related to the relation instance. For instance, sentence a) will be processed into “Two cases of [[amisulpride]] overdose: a cause for << prolonged QT syndrome >>”. In the Results section, we will describe the advantages and disadvantages of the 2 different pretreatment methods.

Figure 2. An instance with two different biomedical entity pretreatments.

Positive instance of amisulpride (C1)-prolonged QT syndrome (D2)
a) Two cases of **chemical overdose**: a cause for **disease**.
b) Two cases of deliberate self-**poisoning** with 5 g and 3.6 g of **chemical**, respectively, are reported.
c) In both cases, **disease** and **hypocalcaemia** were noted.
d) The **chemical** appeared to respond to administration of intravenous **calcium gluconate**.

a) The replacement method

Positive instance of amisulpride (C1)-prolonged QT syndrome (D2)
a) Two cases of **[[amisulpride]] overdose**: a cause for << **prolonged QT syndrome** >>.
b) Two cases of deliberate self-**poisoning** with 5 g and 3.6 g of **[[amisulpride]]**, respectively, are reported.
c) In both cases, < < **QT prolongation** > > and **hypocalcaemia** were noted.
d) The << **QT prolongation** >> appeared to respond to administration of intravenous **calcium gluconate**.

b) The addition method

Model Architecture

As shown in Figure 3, when adopting this data preprocessing, the length of most instances is very long, which results from very long-distance dependencies and complex semantics.

Self-attention structure can directly calculate similarities between words, so that the distance between words is 1, which can intuitively solve long-distance dependencies. As demonstrated by Tang et al [21], Transformer, a combined self-attention structure, is capable of semantic feature extraction far exceeding that of RNN and CNN and performs better when increasing the number of attention structures. Therefore, a pretrained self-attention structure, namely a pretrained transformer, is applied for these problems.

However, for document-level relationship extraction, according to our preprocessing method, the length of instances is longer than the experimental data in the paper by Tang et al [21], and the semantics are more complicated. There are multiple target entity pairs in the instances; some reflect the correct relationship, and some do not. Therefore, the transformer structure must have a certain reasoning ability. To verify the validity of a pretrained self-attention structure on document-level relation extraction, we adopted the structure of SciBERT, which was pretrained on the scientific literature, and added a feed-forward network (FNN) as a classifier. A visual model architecture is provided in Figure 3. We fine-tuned the model on the preprocessed CDR dataset. The structure of model is described in detail in the following paragraphs.

Basde on the structure of BERT, SciBERT built a new vocabulary, called SCIVOCAB, and was trained on a scientific corpus that consists of computer science domain and biomedical papers. Following SciBERT, we still employ the same input representation, constructed by summation of token embedding, segment embedding, and position embedding. The tokens "[CLS]" and "[SEP]" are added at the beginning and end, respectively, of each instance. In addition, when tokenizing words, WordPiece embedding [23] was used with SCIVOCAB to separate words and split word pieces with "##".

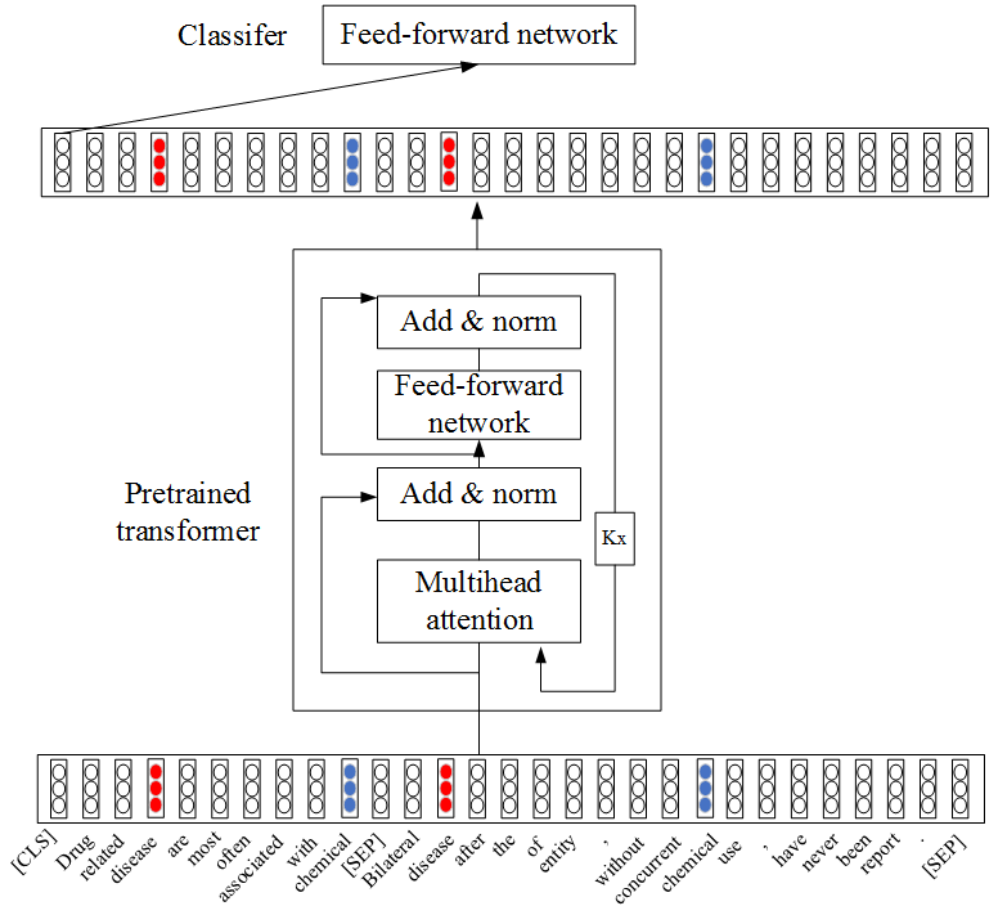
SciBERT is made up of N transformer stacks. Transformer stack k is denoted by $Transformer_k$, which has its own parameters and consists of 2 components: multi-head attention and FFN.

$$Sk = Transformerk(Sk\text{-}1) \quad (1)$$

where $S^k \in R^{n\times d}$ is the output of the transformer stack k. S^0 is the input representation of text sequence X, $X \in R^{n\times d}$. n is the length of text sequence, and d is the dimension of input representation. The whole text sequence shares the same parameters as the transformers.

The multihead applies self-attention, or scaled dot-product attention, multiple times. Through the mapping of the query Q, key K, and value V, scaled dot-product attention obtains a weighted sum of the values. $Q, K, V \in R^{n\times d}$ are the same matrices in the self-attention computation that are the input of transformer.

Figure 3. The architecture of the model.



Instead of applying a single scaled dot-product attention, multihead attention applies query *Q*, key *K*, and value *V* to linearly project the input *h* times with different, learned linear projections to *n* x *l* dimensions, respectively, where *l* = *d*/*h* and *h* is the number of the head. The reason for that is multihead attention can form different representation subspaces at different positions, learn more semantic information, and capture long-distance dependencies better.

$$Oh = \text{softmax}(QWiQ(KWiK)T / \text{sqrt}(dh))VWiV \quad (2)$$

Where the projections are parameter matrices $W^Q \in R^{d\times l}$, $W^K \in R^{d\times l}$, $W^V \in R^{d\times l}$, and $O_h \in R^{dxl}.sqrt(d_h)$ is a scale factor to prevent the result of the dot-product attention from enlarging, and *sqrt( )* indicates that the square root is extracted.

Then, the outputs of the individual attention heads are merged, denoted as $O \in R^{dxl}$. The input and output of the multihead attention are connected by residual connection. Layer normalization, denoted *LN*, is applied to the output of the residual connection.

$$O = [o_1;...;o_h] \quad (3)$$

$$M = LN(S^{k-1}+O) \quad (4)$$

Where $M \in R^{n\times d}$

The second component of the transformer stack is 2 layers of pointwise FFN. On the other hand, it can be described as 2 convolutions with kernel size 1.

$$S^k = ReLU(MW_1+b_1)W_2+b_2 \quad (5)$$

where $W_1 \in R^{d\times m}$, $b_1 \in R^{n\times m}$, $W_2 \in R^{m\times d}$, and $b_2 \in R^{n\times d}$. Each row of or is the same, and $m = 4d$.

The final layer is an FFN, a relation classifier. It corresponds to the final output of transformer of the token "[CLS]".

$$c = W^{pred}s_1 \quad (6)$$

Where $W^{pred} \in R^{o\times d}$ is the weight matrix and s $\in R^d$ is the final output of the token "[CLS]".

Results

Overview

In this section, we first describe some experimental datasets and provide some experiment settings. Then, we compare the performance of SciBERT with that of existing methods and validate the availability of the pretrained self-attention structure on the document-level dataset through the visualization of the multihead attention. Finally, experimenting on different datasets, including 2 sentence-level corpora and a document-level corpus, we compare various biomedical entity pretreatments and analyze which preprocessing is better for the self-attention structure.

Datasets

Table 1 shows the statistics of the CDR [3], protein-protein interactions affected by mutations (PPIm) [24], DDI [1], and CPR [2] datasets. The CDR and PPIm datasets are

document-level annotated corpus, and the DDI and CPR datasets are sentence-level annotated corpora, which are only used to discuss the advantages and disadvantages of different biomedical entity pretreatments.

Table 1. Descriptions of the chemical-disease relation datasets.

Dataset, Types	Training set	Development set	Test set
CDR[a]			
Documents	500	500	500
Positive	1038	1012	1066
Negative	4324	4134	4374
PPIm[b]			
Documents	597	N/A[c]	635
Positive	750	N/A[c]	869
Negative	1401	N/A[c]	1717
DDI[d]			
Sentence	18,872	N/A[c]	3843
Positive	3964	N/A[c]	970
Negative	14,908	N/A[c]	2873
Int	183	N/A[c]	96
Advice	815	N/A[c]	219
Effect	1654	N/A[c]	357
Mechanism	1312	N/A[c]	298
CPR[e]			
Sentences	6437	3558	5744
Positive	4172	2427	3469
Negative	2265	1131	2275
CPR:3	777	552	667
CPR:4	2260	1103	1667
CPR:5	173	116	198
CPR:6	235	199	293
CPR:9	727	457	644

[a]CDR: chemical-disease relation.

[b]PPIm: protein-protein interaction affected by mutations.

[c]Development sets do not exist in the PPIm an dDDI datasets.

[d]DDI: drug-drug interaction.

[e]CPR: chemical-protein relation.

The CDR dataset is used to extract CID and is a 2-label classification task. The PPIm dataset is released to extract protein-protein interactions affected by genetic mutations, which is a 2-label classification. Aimed at extracting drug-drug interactions, DDI is concerned with classifying into 5 relation types, including the int type, advice type, effect type, mechanism type, and negative type. For the DDI dataset, we adopted some rules to filter some negative sentences as described by Quan et al [25]. With the purpose of extracting chemical-protein relations, the CPR dataset is labeled as 10 types of chemical-protein relations, 5 of which are used for evaluation. The chemical-protein relations of CPR are classified into 6 categories.

Due to the size of the CDR dataset, we merged the training and development sets to construct the training set. After preprocessing the CDR and PPIm datasets, we counted the average number of sentences per instance, average number of tokens per instance, and average number of tokens per sentence in the constructed instance set. Table 2 shows the statistics of the constructed instance set.

Experiment Setup

We employed the parameters of the uncased SciBERT with the vocabulary SCIVOCAB and fine-tuned on the CDR datasets. The model parameters are described as: $\text{SciBERT}_{\text{uncased}}$: $k = 12$, $h = 12$, $d = 768$, $m = 3072$.

Due to the distinction of the length of instances in the dataset, the input dimensions of the corresponding model for each dataset are different. For the CDR and PPIm datasets, the length of the input sequence is set to 512, and the batch size is set to 6. For the DDI dataset, the length of the input sequence is set to 150, and the batch size is set to 32. For the CPR dataset, the length of the input sequence is set to 200, and the batch size is set to 23. The epoch of all model training is set to 3. All results are averaged across 5 runs. For consistency of comparisons, we merged the training and development sets to train the models.

Table 2. Statistics of the constructed instance sets of the chemical-disease relation (CDR) and protein-protein interaction affected by mutations (PPIm) datasets.

Dataset, Types	Training set	Test set
CDR with preprocessing		
Instances	10,407	5418
Positive	1947	1042
Negative	8460	4376
Sentences per instance	11.1	12.1
Tokens per instance	161.5	168.9
Tokens per sentence	14.6	14.0
PPIm with preprocessing		
Instances	2151	2586
Positive	750	869
Negative	1401	1717
Sentences per instance	9.0	8.8
Tokens per instance	169.6	186.6
Tokens per sentence	18.7	21.2

Comparison of the Pretrained Self-Attention Structure With Other Methods

For the CDR dataset, we compared our method with 6 state-of-the-art models without any knowledge bases. Zhou et al [9] proposed a method based on feature engineering and long short-term memory. Gu et al [8] combined CNN with maximum entropy. A recurrent piecewise CNN [13] was the piecewise CNN. A bi-affine relation attention network [14] incorporated an attention network, multi-instance learning, and multitask learning. A labeled edge graph CNN [12] was used for document-level dependency graphs. For the PPIm dataset, we compared our method with 4 models. Because few studies focused on the PPIm dataset, the 4 models are not really state-of-the-art. Table 3 shows the result of the comparisons.

As shown in Table 3, compared with other approaches, our method with the replacement method greatly improved the precision. The F1 score is 1.9% higher than the best result from Vargas et al [19] with the CDR test set. Our method also has great performance with the PPIm. It shows that a pretrained self-attention structure can be suitable for a document-level dataset.

Table 3. Performance of the chemical-disease relation (CDR) and protein-protein interactions affected by mutations (PPIm) test datasets compared with state-of-the-art methods.

Dataset, Model	P[a], %	R[b], %	F1, %
CDR			
LSTM[c] [9]	55.6	68.4	61.3
CNN[d] [8]	55.7	68.1	61.3
RPCNN[e] [13]	55.2	63.6	59.1
BRAN[f] [14]	55.6	70.8	62.1
GCNN[g] [12]	52.8	66.0	58.6
Our method	65.5	62.6	64.0
PPIm			
SVM[h] [26]	32.0	34.0	33.0
CNN (without KB[i]) [27]	38.2	37.3	37.8
MNM[j] [28]	40.3	32.3	35.9
MNM+Rule [28]	38.0	37.0	37.5
Our method	83.5	90.4	86.8

[a]P: precision.

[b]CNN: convolutional neural network.

[c]R: recall.

[d]LSTM: long short-term memory.

[e]RPCNN: recurrent piecewise convolutional neural network.

[f]BRAN: bi-affine relation attention network.

[g]GCNN: graph convolutional neural network.

[h]SVM: support vector machine.

[i]KB: knowledge base.

[j]MNM: memory neural network.

Effects of Pretreatment Methods for Biomedical Entities

As described later, there are 2 methods, one of which is the replacement method, that replace biomedical entities with uniform words. The second method is the addition method, which adds extra tags in the left and right sides of biomedical entities. We conducted experiments with the CDR, PPIm, DDI, and CPR datasets. The comparison of the 2 pretreatments for biomedical entities is shown in Table 4.

For each dataset, the recall rate and F1 score obtained with our model with the replacement method were higher than obtained with our model with the addition method, especially for the CDR dataset. The reason is that biomedical entities are complicated, and most are compound words. For the pretrained self-attention structure, the word embeddings of biomedical entities are hard to learn from small biomedical datasets. As a consequence, replacing the target entities with uniform words is beneficial for the model to understand target entities and pay more attention in the context of target entities.

Table 4. Comparison of 2 pretreatments (addition and replacement) for biomedical entities using our method.

Dataset, Types	Addition method			Replacement method		
	P[a], %	R[b], %	F1, %	P, %	R, %	F1, %
CDR[c]						
Positive	67.4	54.8	60.4	65.5	62.6	64.0
PPIm[d]						
Positive	79.3	91.5	84.8	83.5	90.4	86.8
DDI[e]						
Int	74.8	46.2	57.1	76.2	46.9	58.0
Advise	87.2	84.7	85.9	88.6	89.0	88.8
Effect	77.2	82.1	79.5	77.0	82.6	79.7
Mechanism	84.8	80.4	82.5	82.1	86.0	84.0
All	81.6	78.6	80.0	81.2	81.3	81.4
CPR[f]						
CPR:3	73.5	80.3	76.7	75.4	79.5	77.4
CPR:4	84.4	88.8	86.6	83.7	90.4	86.9
CPR:5	80.7	82.0	81.3	81.2	86.5	83.7
CPR:6	84.0	89.4	86.7	86.5	88.2	87.3
CPR:9	76.2	86.9	81.2	79.5	90.1	84.5
All	80.4	86.5	83.3	81.4	87.9	84.5

[a]P: precision.

[b]R: recall.

[c]CDR: chemical-drug reaction.

[d]PPIm: protein-protein interaction affected by mutations.

[e]DDI: drug-drug interaction.

[f]CPR: chemical-protein reaction.

Comparison of Different Pretrained Models

BERT and SciBERT are the pretrained models that have the same self-attention structure. The difference between the two is that BERT is pretrained on the wiki corpus and SciBERT is pretrained on a large quantity of scientific papers from the computer science and biomedical domains. Table 5 presents the comparison of BERT and SciBERT on 4 biomedical data sets. As shown by Table 5, SciBERT performs better than BERT, particularly with the F1 score, which was improved by 3.5% on the CDR data set. Therefore, the model pretrained on the biomedical corpus is beneficial for extracting biomedical relations.

Table 5. Comparison of different pretrained models using our method.

Dataset, Type	BERT			SciBERT		
	P[a], %	R[b], %	F1, %	P, %	R, %	F1, %
CDR[c]						
Positive	62.9	58.3	60.5	65.5	62.6	64.0
PPIm[d]						
Positive	79.0	92.2	85.1	83.5	90.4	86.8
DDI[e]						
Int	69.8	42.7	52.8	76.2	46.9	58.0
Advise	91.3	89.0	90.1	88.6	89.0	88.8
Effect	74.1	77.6	75.7	77.0	82.6	79.7
Mechanism	78.5	80.1	79.3	82.1	86.0	84.0
All	79.0	77.5	78.2	81.2	81.3	81.4
CPR[f]						
CPR:3	73.8	76.5	75.1	75.4	79.5	77.4
CPR:4	81.7	89.7	85.5	83.7	90.4	86.9
CPR:5	79.3	80.7	79.9	81.2	86.5	83.7
CPR:6	80.2	84.6	82.2	86.5	88.2	87.3
CPR:9	76.5	88.2	81.9	79.5	90.1	84.5
All	79.0	85.9	82.3	81.4	87.9	84.5

[a]P: precision.

[b]R: recall.

[c]CDR: chemical-drug reaction.

[d]PPIm: protein-protein interaction affected by mutations.

[e]DDI: drug-drug interaction.

[f]CPR: chemical-protein reaction.

Analysis of Each Component of the Method

Data preprocessing (DP) and pretraining means (PTM) are important components of our method; DP aims to alleviate noise, and PTM is designed to solve the long-distance dependencies. We compared the importance of each component of our method with the CDR dataset. Table 6 shows the changes in performance on the CDR dataset by removing DP and PRM. PTM resulted in a greater performance improvement than DP.

Table 6. Performance changes by removing different parts of our model.

CDR[a] dataset	P[b], %	R[c], %	F1, %	Change, %
Baseline	65.5	62.6	64.0	N/A
Remove DP[d]	67.0	54.3	60.0	–6.3
Remove PTM[e]	46.1	39.5	42.6	–33.4
Remove DP and PTM	48.9	31.2	38.1	–40.5

[a]CDR: chemical-drug reaction.

[b]P: precision.

[c]R: recall.

[d]DP: data preprocessing.

[e]PTM: pertraining means.

Discussion

Principal Findings

To fully illustrate that our model can solve the problem of long-distance dependencies, we set 50 as the unit of instance length to count the number of positive and negative instances of the CDR test set, as shown in Table 7. As can be seen from the table, the instance length of the test sets is concentrated in the range of 50 to 300.

We calculated the precision rate, recall rate, and accuracy rate of each interval length in the test set. The results are shown in Table 8. As can be seen in the table, the model has good performance when the instance length is longer than 100, except for the instances with lengths of 201 to 250. Therefore, our model can capture long-distance dependencies.

Table 7. Quantity distribution of the chemical-disease relation test set.

Interval length	Positive	Negative	Sum
0-50	70	344	414
51-100	181	884	1065
101-150	200	845	1045
151-200	160	756	916
201-250	158	663	821
251-300	177	571	748
301-350	56	216	272
351-400	34	64	98
>400	6	33	39

Table 8. Results of each interval length in the test set using our replacement method.

Interval length	P[a], %	R[b], %	F1, %
0-50	54.2	64.3	58.8
51-100	57.5	57.5	57.5
101-150	67.7	64.0	65.8
151-200	64.4	71.2	67.7
201-250	66.2	54.4	59.7
251-300	69.9	69.5	69.7
301-350	70.8	60.7	65.4
351-400	80.0	82.4	81.2
>400	100.0	66.7	80.0

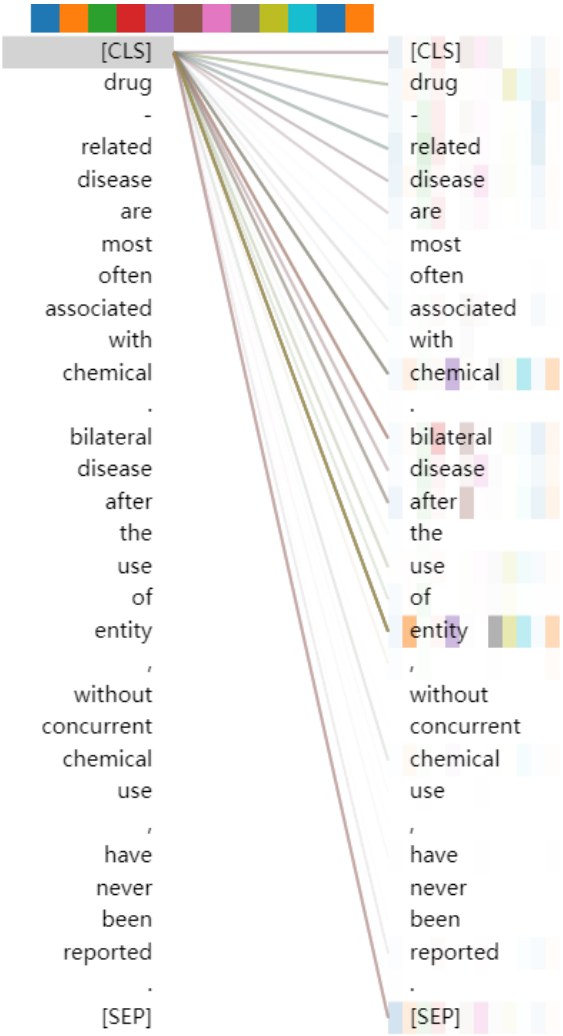
[a]P: precision.

[b]R: recall.

To verify that the pretrained self-attention mechanism works as we believe, which is that it can take advantage of the textual context and capture very long-range dependencies to understand the complex semantics of biomedical text, we visualized the output of token "[CLS]" in the multihead of the final transformer stack, as shown in Figure 4.

As seen by the token colors, the token "[CLS]" is related to the following tokens: "chemical," "disease," "drug," "related," "bilateral," "[CLS]," and "[SEP]". The 12 different colors refer to the different head attentions. The more and darker the colors, the more relevant the token. Lines between two tokens denote a correlation between two tokens. Their clarity depends on the result of the head attentions.

Figure 4. Visualization of the output of token "[CLS]" in the multihead attention of the final transformer stack.



From the perspective of semantic analysis, there are 2 places in this example reflecting a relationship between a disease and chemical: "Drug-related disease are most often associated with chemical." and "Bilateral disease after the use of entity, without concurrent chemical use, have never been reported." In the first sentence, the relation between chemical and disease is mainly determined by the following tokens: "associated," "chemical," "disease," "related," and "drug." In the second sentence, the relation between chemical and disease is mainly determined by the following tokens: "reported," "never," "chemical," "disease," "without," and "concurrent." Token "[CLS]" is related to the most keywords in both sentences. Therefore, the pretrained self-attention structure can take advantage of the textual context and capture very long-range dependencies from document-level instances. On the other hand, the distribution of the different colors shows that multihead attention can form diverse representation subspaces to learn more complicated semantics.

However, from the gradation of the colors, the relationship between token "[CLS]" and the keywords is not strong enough. Token "[CLS]" is not highly correlated with token "disease" in this instance. We visualized the output of tokens "chemical" and "disease" in the final multihead attention, as shown in Figures 5 and 6. As seen in these figures, the tokens "chemical" and "disease" in the sentences capture more local information, compared with the token "[CLS]." It may be inferred that, for document-level relation extraction in the final layer of the pretrained self-attention structure, designing a special network to capture the relationships between different target entities is better than applying a dense layer.

Figure 5. Visualization of the output of token “disease” in the final multihead attention.

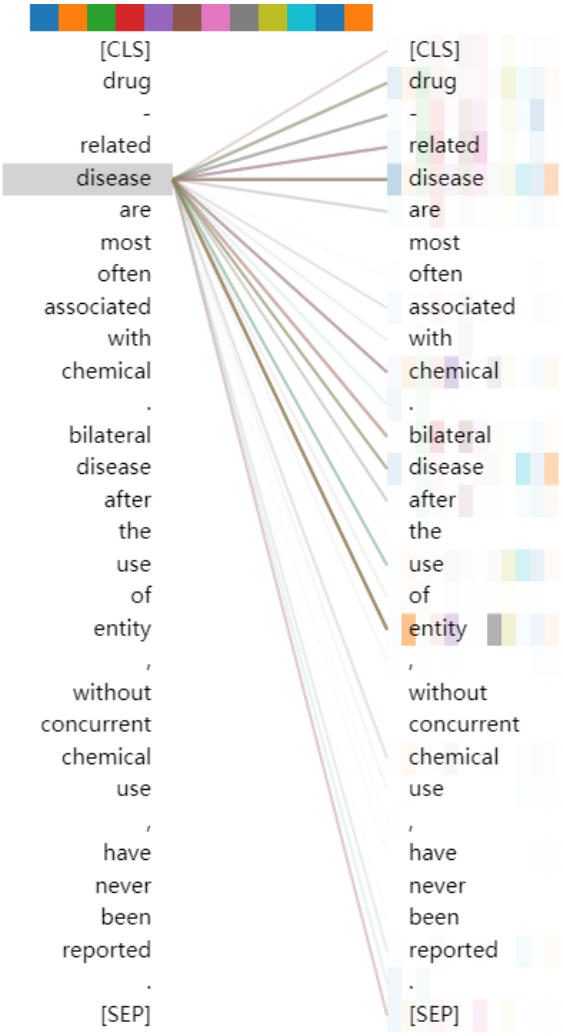
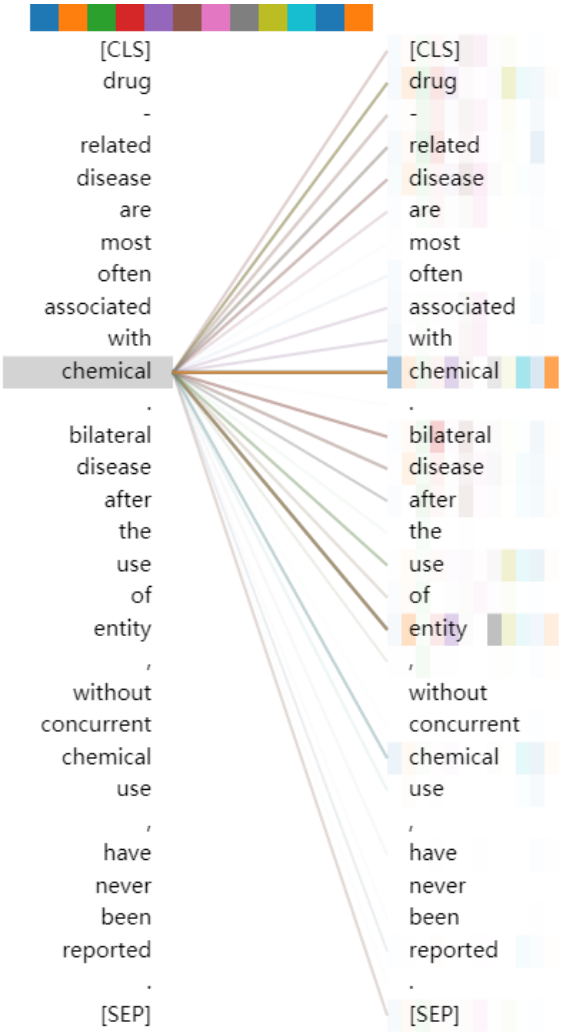


Figure 6. Visualization of the output of token “chemical” in the final multihead attention.



Conclusions

For a document-level annotated dataset, instead of dividing the dataset, we considered all target entity pairs as a whole and applied a pretrained self-attention structure to extract biomedical relations. The results and analysis show that the pretrained self-attention structure extracted relations of multiple entity pairs in a document. Through the visualization of the transformer, we verified that the pretrained self-attention structure can capture long-distance dependencies and learn complicated semantics. Furthermore, we conclude that replacement of biomedical entities benefits biomedical relation extraction, especially for document-level relation extraction.

However, this method still has some issues. In future work, we plan to design a more effective network to capture local relations between biomedical entities and improve our method.

Acknowledgments

This study was funded by the Natural Science Foundation of Guangdong Province of China (2015A030308017), National Natural Science Foundation of China (61976239), and Innovation Foundation of High-end Scientific Research Institutions of Zhongshan City of China (2019AG031).

Conflicts of Interest

None declared.

References

1. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014 Jan;42(Database issue):D1091-D1097 [FREE Full text] [doi: 10.1093/nar/gkt1068] [Medline: 24203711]

2. Krallinger M, Rabal O, Akhondi S. Overview of the BioCreative VI chemical-protein interaction Track. 2017 Oct Presented at: In: Proceedings of the sixth BioCreative challenge evaluation workshop; 2017; Bethesda, Maryland p. 141-146.
3. Li J, Sun Y, Johnson RJ, Sciaky D, Wei C, Leaman R, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw068] [Medline: 27161011]
4. Zhang Y, Zheng W, Lin H, Wang J, Yang Z, Dumontier M. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics 2018 Mar 01;34(5):828-835 [FREE Full text] [doi: 10.1093/bioinformatics/btx659] [Medline: 29077847]
5. Sun C, Yang Z, Wang L. Hierarchical Recurrent Convolutional Neural Network for Chemical-protein Relation Extraction from Biomedical Literature. 2018 Mar Presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 2018; Madrid, Spain p. 3-6. [doi: 10.1109/bibm.2018.8621159]
6. Lim S, Kang J. Chemical-gene relation extraction using recursive neural network. Database (Oxford) 2018 Jan 01;2018 [FREE Full text] [doi: 10.1093/database/bay060] [Medline: 29961818]
7. Peng Y, Rios A, Kavuluru R, Lu Z. Extracting chemical-protein relations with ensembles of SVM and deep learning models. Database (Oxford) 2018 Jan 01;2018 [FREE Full text] [doi: 10.1093/database/bay073] [Medline: 30020437]
8. Gu J, Sun F, Qian L, Zhou G. Chemical-induced disease relation extraction via convolutional neural network. Database (Oxford) 2017 Jan 01;2017(1) [FREE Full text] [doi: 10.1093/database/bax024] [Medline: 28415073]
9. Zhou H, Deng H, Chen L, Yang Y, Jia C, Huang D. Exploiting syntactic and semantics information for chemical-disease relation extraction. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw048] [Medline: 27081156]
10. Peng NY, Hoifung P, Chris Q. Cross-sentence n-ary relation extraction with Graph LSTMs. Transactions of the Association for Computational Linguistics ? 2017;5:101-115.
11. Song L, Zhang Y, Wang Z, Gildea D. N-ary relation extraction using graph-state LSTM. 2018 Oct Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium.
12. Sunil K, Fenia C, Makoto M. Inter-sentence relation extraction with document-level graph convolutional neural network. 2019 Jun Presented at: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; July; Florence, Italy; 2019; Italy. [doi: 10.18653/v1/p19-1423]
13. Li H, Yang M, Chen Q, Tang B, Wang X, Yan J. Chemical-induced disease extraction via recurrent piecewise convolutional neural networks. BMC Med Inform Decis Mak 2018 Jul 23;18(Suppl 2):60 [FREE Full text] [doi: 10.1186/s12911-018-0629-3] [Medline: 30066652]
14. Verga P, Strubell E, McCallum A. Simultaneously Self-Attending to All Mentions for Full-Abstract Biological Relation Extraction. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers).: Association for Computational Linguistics; 2018 Presented at: 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; June 6, 2018; New Orleans, LA p. 872-884. [doi: 10.18653/v1/n18-1080]
15. Wei C, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw032] [Medline: 26994911]
16. Gu J, Qian L, Zhou G. Chemical-induced disease relation extraction with various linguistic features. Database (Oxford) 2016;2016 [FREE Full text] [doi: 10.1093/database/baw042] [Medline: 27052618]
17. Panyam NC, Verspoor K, Cohn T, Ramamohanarao K. Exploiting graph kernels for high performance biomedical relation extraction. J Biomed Semantics 2018 Jan 30;9(1):7 [FREE Full text] [doi: 10.1186/s13326-017-0168-3] [Medline: 29382397]
18. Stratas NE. A double-blind study of the efficacy and safety of dothiepin hydrochloride in the treatment of major depressive disorder. J Clin Psychiatry 1984 Nov;45(11):466-469. [Medline: 6386793]
19. Beltagy I, Lo K, Cohan A. SciBERT: A Pre-trained Language Model for scientific text. SciBERT; 2019 Nov Presented at: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing; 2019; Hong Kong. [doi: 10.18653/v1/d19-1371]
20. Devlin J, Chang M, Lee K. BERT: Pre-training of deep bidirectional transformers for language understanding. Computation and Language 2019 May:2019.
21. Tang G, Müller M, Rios A, Sennrich R. Why self-attention? a targeted evaluation of neural machine translation architectures. 2018 Nov Presented at: Conference on Empirical Methods in Natural Language Processing; 2018; Brussels, Belgium.
22. Ward DI. Two cases of amisulpride overdose: a cause for prolonged QT syndrome. Emerg Med Australas 2005 Jun;17(3):274-276. [doi: 10.1111/j.1742-6723.2005.00734.x] [Medline: 15953230]
23. Wu Y, Mike S, Chen Z. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv.08144 2016 Oct:2016.
24. Dogan R, Chatr-aryamontri A. BioCreative VI Precision Medicine Track: creating a training corpus for mining protein-protein interactions affected by mutations. 2017 Aug Presented at: =BioNLP 2017 workshop; 2017; Vancouver, Canada. [doi: 10.18653/v1/w17-2321]
25. Quan C, Hua L, Sun X, Bai W. Multichannel Convolutional Neural Network for Biological Relation Extraction. Biomed Res Int 2016;2016:1850404 [FREE Full text] [doi: 10.1155/2016/1850404] [Medline: 28053977]

26. Fan Z, Soldaini L, Cohan A, Goharian N. Relation Extraction for Protein-protein Interactions Affected by Mutations. 2018 Aug Presented at: BCB '18th ACM International Conference on Bioinformatics, Computational Biology Health Informatics; 2018; Washington DC USA. [doi: 10.1145/3233547.3233617]
27. Tran T, Kavuluru R. An end-to-end deep learning architecture for extracting protein-protein interactions affected by genetic mutations. Database (Oxford) 2018 Jan 01;2018:1-13 [FREE Full text] [doi: 10.1093/database/bay092] [Medline: 30239680]
28. Zhou H, Liu Z, Ning S, Yang Y, Lang C, Lin Y, et al. Leveraging prior knowledge for protein-protein interaction extraction with memory network. Database (Oxford) 2018 Jan 01;2018 [FREE Full text] [doi: 10.1093/database/bay071] [Medline: 30010731]

Abbreviations

BRAN: bi-affine relation attention network.
CDR: chemical-disease relation.
CID: chemical-induced disease.
CNN: convolutional neural network.
CPR: chemical-protein relation.
DDI: drug-drug interaction.
DP: data preprocessing.
FNN: feed-forward network.
GCNN: graph convolutional neural network.
KB: knowledge base.
MNM: memory neural network.
P: precision.
PPIm: protein-protein interactions affected by mutations.
PTM: pretraining means.
R: recall.
RNN: recurrent neural network.
RPCNN: recurrent piecewise convolutional neural network.
SVM: support vector machine.

Edited by T Hao, B Tang, Z Huang; submitted 30.12.19; peer-reviewed by Z Zhao; comments to author 14.02.20; revised version received 02.03.20; accepted 19.03.20; published 29.05.20.

Please cite as:
Liu X, Fan J, Dong S
Document-Level Biomedical Relation Extraction Leveraging Pretrained Self-Attention Structure and Entity Replacement: Algorithm and Pretreatment Method Validation Study
JMIR Med Inform 2020;8(5):e17644
URL: http://medinform.jmir.org/2020/5/e17644/
doi:10.2196/17644
PMID:32469325

Original Paper

Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study

Yihao Huang[1,2], MD; Mingtao Li[3], MD, PhD

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China

[2]Complex Systems Research Center, Shanxi University, Taiyuan, China

[3]College of Mathematics, Shanxi University of Technology, Taiyuan, China

Corresponding Author:
Yihao Huang, MD
School of Computer and Information Technology
Shanxi University
92 Wucheng Road
Taiyuan, 030006
China
Phone: 86 15834136789
Email: 297535248@qq.com

Abstract

Background: Rabies is an acute infectious disease of the central nervous system caused by the rabies virus. The mortality rate of rabies is almost 100%. For some countries with poor sanitation, the spread of rabies among dogs is very serious.

Objective: The objective of this paper was to study the ecological transmission mode of rabies to make theoretical contributions to the suppression of rabies in China.

Methods: A mathematical model of the transmission mode of rabies was constructed using relevant data from the literature and officially published figures in China. Using this model, we fitted the data of the number of patients with rabies and predicted the future number of patients with rabies. In addition, we studied the effectiveness of different rabies suppression measures.

Results: The results of the study indicated that the number of people infected with rabies will rise in the first stage, and then decrease. The model forecasted that in about 10 years, the number of rabies cases will be controlled within a relatively stable range. According to the prediction results of the model reported in this paper, the number of rabies cases will eventually plateau at approximately 500 people every year. Relatively effective rabies suppression measures include controlling the birth rate of domestic and wild dogs as well as increasing the level of rabies immunity in domestic dogs.

Conclusions: The basic reproductive number of rabies in China is still greater than 1. That is, China currently has insufficient measures to control rabies. The research on the transmission mode of rabies and control measures in this paper can provide theoretical support for rabies control in China.

(JMIR Med Inform 2020;8(5):e18627) doi:10.2196/18627

KEYWORDS

rabies; computer model; suppression measures; basic reproductive number

Introduction

Since ancient times, infectious diseases have brought much distress to human beings, and recorded deaths from infectious diseases abound. For example, the Black Death that killed a quarter of the European population during the medieval period was transmitted by fleas and black rats. In the mid-14th century, a plague killed another 25 million people [1]. The number of diseases transmitted from animals to humans has been increasing. By mutating, viruses become better adapted to their environment and the physiological status of various hosts [2]. More than 6% of infectious diseases that affect human beings originate from animals, and more than half of animal diseases can be transmitted to humans; as such, researchers pay particular attention to such diseases [3]. The zoonotic disease studied in this paper is rabies, which is caused by the rabies virus and can damage the central nervous system in humans. Rabies has a fatality rate of 100%, the highest in the world [4]. Although the fatality rate of rabies is extremely high, effective control measures can prevent the transmission of the disease.

Rabies is a zoonotic acute infectious disease of the central nervous system caused by the rabies virus. As patients with rabies have the clinical manifestation of being afraid of drinking water, this disease was previously called hydrophobia. However, animals with rabies do not have this characteristic [5]. The main symptoms are mania, anxiety, fear of wind and water, salivation, and pharyngeal muscle spasm. Paralysis is a life-endangering symptom. One of the characteristics of this disease is that different patients have incubation periods of different lengths. Most cases occur within 3 months of infection, with 4%-10% occurring after more than 6 months, and about 1% of cases exceeding 1 year. The longest incubation time reported in the literature is 10 years [6]. Factors that affect the length of the incubation period are age (shorter in children), wound site (shorter if on the head or face), depth of wound (shorter with deeper wounds), the amount of virus, the virulence of the strain, and whether formalized debridement processing and preventive vaccination were performed. Factors such as trauma, cold, and overwork may contribute to early onset [7].

Humans have been aware of the disease rabies since the time of Ancient Babylonians. At that time, there were no effective rabies control methods. Due to the pain and high fatality rate of rabies, many people committed suicide after being bitten by dogs [8]. Today, there are more than 150 countries in the world that have a large number of deaths contributed to the wild spread of rabies. At present, only one-quarter of the world's countries effectively control the occurrence of rabies. In some countries where severe health conditions are prevalent, the number of rabies cases is high and the spread is difficult to control. In Asia, India has the most severe rabies situation, with an annual incidence of 20,000 cases, followed by Southeast Asia and most parts of Africa.

In China, the vaccination rate of dogs is relatively low, especially in rural areas, where free-range dogs are rarely vaccinated. As such, the death toll from rabies is significant. Over the past two decades, the number of people who died of rabies in China first increased and then decreased. The number of deaths from rabies rose rapidly in the 1990s, reaching a peak in 2007. After the unremitting efforts of relevant personnel in China, the number of deaths from rabies began to decrease; deaths had decreased by about 76% as of 2013 [9]. Wild dogs are a source of rabies infections that cannot be ignored as they are not domesticated and may bite humans. Therefore, we conducted a comprehensive study of domestic dogs, wild dogs, and humans in the context of rabies transmission.

Due to the highly infectious nature of the rabies virus, it is being studied by researchers worldwide. Simulation research is a research method that can transform complex biological phenomena into mathematical problems; such mathematical methods may enable better logical reasoning and disease transmission analysis. Some researchers have established mathematical models of rabies transmission from dogs to other animals and studied the transmission mode of rabies in animals to ascertain the optimal time for vaccine control [10]. Other researchers have designed a rabies transmission model based on transmission in the domestic population. In building the model described in this paper, dogs and humans were classified into the following 4 categories: susceptibility, latency, infection, and recovery. The transmission mode of rabies among populations was determined and effective rabies prevention methods were proposed [11]. Other researchers previously designed a transmission model for domestic and wild dogs. Studies have shown that to control the spread of rabies, the management of domestic dog breeding must be strengthened [12]. Therefore, we conducted a simulation study of the ecological infection model of rabies to provide theoretical support for rabies research.

To study the transmission mode of rabies and determine effective measures for suppressing rabies, we first constructed a mathematical model for the transmission mode of rabies. To construct the mathematical model, data from the literature and officially published data in China was used. The model fits the number of patients with rabies and predicts the number of patients with rabies in the future. In addition, we studied the effectiveness of different rabies suppression measures. This paper will provide theoretical support for rabies suppression in China.

Methods

Establishment of a Mathematical Model of Rabies Transmission

Rabies transmission models of humans, wild dogs, and domestic dogs were established. Dogs and humans were divided into 4 categories: susceptibility, latency, infection, and recovery. S_a, E_a, I_a, and R_a represent the 4 categories of susceptibility, latency, infection, and recovery in wild dogs. S_b, E_b, I_b, and R_b indicate the 4 categories of susceptibility, latency, infection, and recovery in domestic dogs. S_c, E_c, I_c, and R_c represent the 4 categories of susceptibility, latency, infection, and recovery in humans. Before the simulation study, it was hypothesized that only dogs that have been in contact with an infected person would become infected and dogs in the incubation period have both vaccination and autoimmunity. The transmission model of rabies is shown in Figure 1. The arrows in the figure represent the direction of rabies transmission.

According to this, the corresponding mathematical model was established. The mathematical models for transmission among dogs are shown in Equations 1-8.

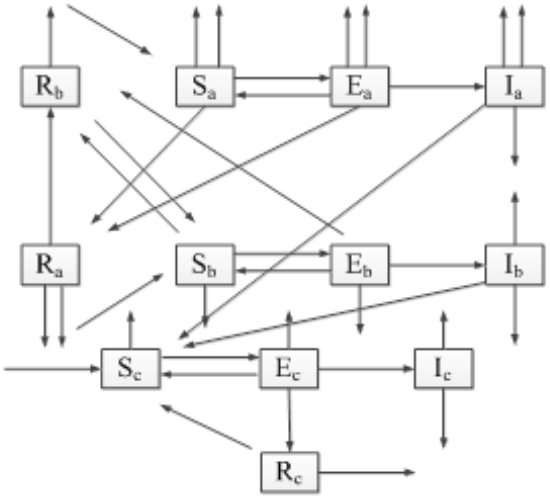
The mathematical models for human transmission are shown in Equations 9-12.

Corresponding to the actual situation, the parameters in Equations 1-12 are nonnegative.

For both dog groups, A indicates the birth rate and p indicates the ratio of the two dog groups that are classified as susceptible in the incubation period of rabies but that have not developed disease.

$b_aS_aI_a$ and $b_a'S_aI_b$ indicate the number of rabies infections in wild dogs in two dog groups within a unit of time. $b_bS_bI_b$ and $b_b'S_bI_a$ indicate the number of rabies infections in domestic dogs in two dog groups within a unit of time. Additionally, γ indicates the rate of immunity of the two dog groups, μ indicates the natural mortality rate of the two dog groups, α indicates the fatality rate of the two dog groups, λ indicates the failure rate of vaccination in dogs, c_a indicates the killing rate of wild dogs, c_a' indicates the killing rate of immune wild dogs, σ indicates the removal rate between the two dog groups in the incubation and infected periods.

Figure 1. Rabies transmission mode.



Model Dynamics

Model dynamics is used to study the relationship between rabies suppression measures and the number of patients with rabies. For dog groups, it is obtained as follows.

According to the actual situation, the number of groups in this mathematical model is always greater than 0. Then, the equation is obtained as shown below.

Thus, α can obtain the positive invariant set X of the model as follows.

Similarly, the positive invariant set X of the human situation can be obtained as follows:

For the human population, H indicates the birth rate in the human population, p indicates the ratio of the human population that is classified as susceptible during the incubation period of rabies but has not developed disease, and λ_c indicates the failure rate of vaccination in the human population. $b_cS_cI_a$ and $b_c'S_cI_b$ indicate the number of rabies infections in humans caused by the two dog groups within a unit time, γ indicates the rate of immunity of the human population, m indicates the natural mortality rate of humans, and α indicates the fatality rate of the human population.

The rabies-free equilibrium point  in a dog group can be found using Equations 1-8. The rabies-free equilibrium point  in a human population can be found using Equations 9-12. Therefore, the basic reproductive number (BRN) of this rabies transmission mode is shown below.

In dog groups, the BRN is  . The occurrence of cross infection greatly affects the BRN. Therefore, cross infection should be considered in regard to daily protection.

Parameter Estimation

There is currently no fixed database of rabies in China. Therefore, the data in this paper were obtained from previously published literature and reports. The number of patients with rabies used for this paper corresponds to the data reported by the Public Health Scientific Data Center and the National Health and Family Planning Commission. Taking the year as the unit, the data shown in Table 1 was obtained.

Table 1. Data list of parameters used for Equations 1-12.

Parameter	Symbol	Value
Average number of wild dogs born each year	A_a	2.47×10^{5}
Infection rate of susceptible dogs by wild dogs in unit time	β_a	2.91×10^{-6}
Infection rate of susceptible dogs by domestic dogs per unit time	β_b	2.20×10^{-7}
Infection rate of susceptible human by wild dogs in unit time	β_c	3.39×10^{-12}
Killing rate of wild dogs	c_a	0.06
Migration rate from latent wild dogs to infected dogs	σ_a	0.35
Failure rate of vaccination in dogs	λ	0.5
Average number of domestic dog births per year	A_b	1.80×10^{6}
Infection rate of susceptible wild dogs by domestic dogs in unit time	β_a,	6.01×10^{-7}
Infection rate of susceptible domestic dogs by wild dogs in unit time	β_b,	5.01×10^{-7}
Infection rate of susceptible humans by domestic dogs in unit time	β_c,	6.79×10^{-12}
Killing rate of domestic dogs	c_a,	0.4
Immunity rate of wild dogs	γ_a	0.05
Migration rate of humans from incubation period to infection period	σ_c	0.33
Birth rate in the human population	H	1.49×10^{7}
The ratio of wild dogs and domestic dogs during the incubation period that did not have rabies outbreak and recovered to susceptible persons	p_a	0.35
The ratio of domestic dogs in incubation period who did not have rabies outbreak and recovered to susceptible population	p_b	0.37
The ratio of people in incubation period who did not have rabies outbreak and recovered to susceptible population	p_c	0.33
Natural mortality of wild dogs	μ_a	0.24
Disease-caused mortality of wild dogs	α_a	1
Failure rate of vaccination in people	λ_c	1

Results

Fitting and Prediction Results of the Number of Infected Patients

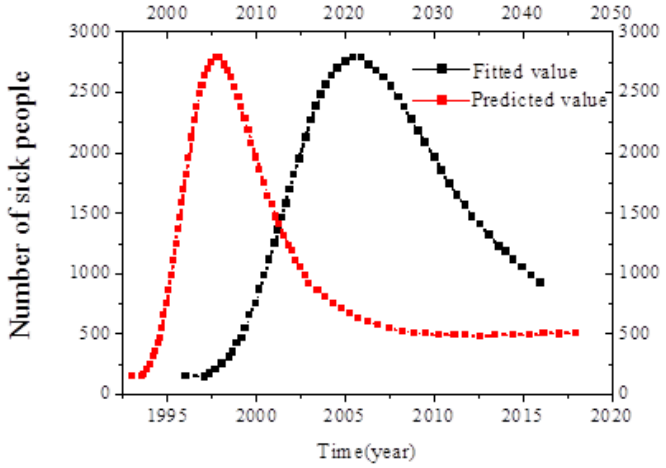
According to the number of patients with rabies in China in the past 10 years, the rabies disease data in the model were fitted. The initial values of each parameter are detailed below.

The initial values of susceptible individuals in the wild dog group, the domestic dog group, and the human population were 2×10^{6}, 3×10^{7}, and 1.29×10^{9}, respectively. The initial values of the latent individuals in the wild dog group, the domestic dog group, and the human population were 7×10^{4}, 2×10^{5}, and 400, respectively. The initial values of infected individuals in the wild dog group, the domestic dog group, and the human population were 2×10^{4}, 5×10^{4}, and 158, respectively. The initial values of recovered individuals in the wild dog group, the domestic dog group, and the human population were 1×10^{5}, 5×10^{6}, and 2×10^{5}, respectively.

Fitting of the model image was performed. During the image fitting process, other unknown parameters were obtained. Figure 2 shows the predicted future number of patients with rabies based on this model.

It can be seen from the above figure that the number of people with rabies rises at first and then decreases. From the fitted data, the number of patients with rabies peaked around 2005. As people's awareness of rabies increases, more people will be vaccinated after being injured by a dog, which will reduce the incidence of rabies and the number of cases. Additionally, due to the development of modern medicine, the effectiveness of the rabies vaccine is gradually increasing, which greatly reduces the number of patients with rabies. Compared to the number of rabies cases in 2016, the number of infected patients is undergoing a continuous decline. In about 10 years, the number of rabies cases is forecasted to be controlled within a relatively stable range. According to the prediction results of this model, the number of rabies cases will eventually be controlled to about 500 people every year. According to the values of the parameters in the model, the BRN is estimated to be 1.069.

Figure 2. Fitting and prediction data of the number of infected patients.



Effect Results of Each Parameter on the BRN

The effect of the birth number of wild (A_a) and domestic dogs (A_b) on the BRN is shown in Figure 3.

It can be seen from the figure above that when more wild dogs are killed, fewer wild dogs are born. Therefore, the value of the BRN can be reduced to less than 1 by killing wild dogs. A reduction in the BRN means that the number of people with rabies will also decrease. Therefore, by killing wild dogs, the incidence of rabies can be effectively reduced. Due to the Chinese people's love of dogs and the tradition of having dogs as pets, the number of domestic dogs is far higher than the number of wild dogs. Therefore, the management of domestic dogs in China is also challenging. As can be seen from the figure above, by controlling the birth number of domestic dogs, the value of the BRN can also be effectively controlled. Thus, the management of domestic dogs in China needs to be strengthened to a certain extent, especially the birth number of domestic dogs, which requires strict control.

The effect of the degree of immunity of wild (γ_a) and domestic (γ_b) dogs on the value of the BRN is shown in Figure 4.

It can be seen in Figure 4 that although an increase in the degree of immunity in wild dogs can reduce the BRN, its effect on the BRN is small and it cannot reduce the BRN to less than 1. This may be because wild dogs are friendly or fearful of humans and generally do not attack humans. Therefore, the degree of immunity in wild dogs does not play a significant role in reducing the number of people with rabies. Conversely, domestic dogs spend a lot of time with people. When playing, it is possible that they will accidentally injure people if their strength is not well-controlled. Thus, for domestic dogs, the degree of immunity can greatly affect the number of people with rabies. Improving the degree of immunity of domestic dogs can greatly reduce the BRN to less than 1.

The values in Figure 4 also show that increasing the immunity of wild dogs can reduce the number of patients with rabies, but it cannot achieve the effect of eliminating rabies. Increasing the immunity of domestic dogs is a practical and effective way to reduce the number of patients with rabies.

Figure 3. The effect of the birth number of wild (A_a) and domestic dogs (A_b) on the R0.

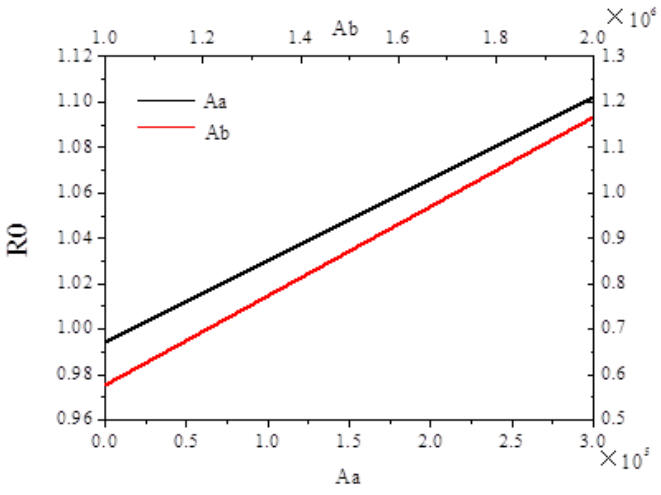
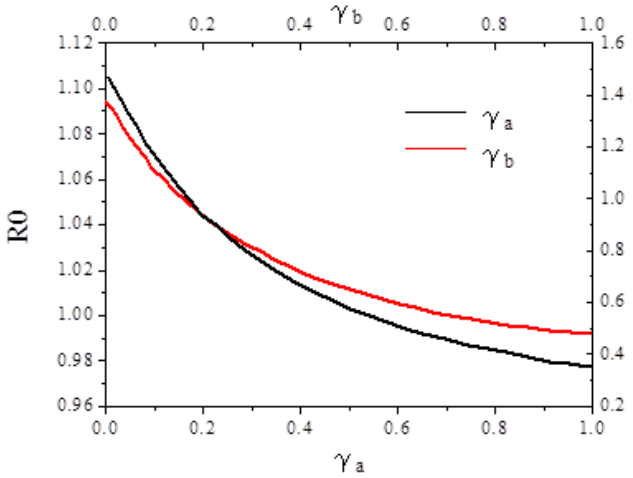


Figure 4. The effect of the degree of immunity of wild (γ_a) and domestic (γ_b) dogs on R0.



Discussion

Overview

Rabies is caused by the rabies virus and can cause damage to the human central nervous system, with a fatality rate of 100%. Among infectious diseases, rabies has the highest fatality rate in the world [13]. Although the fatality rate of rabies is extremely high, the disease can be prevented if effective control measures are taken. The number of patients with rabies in China has declined in recent years, but the number of cases is still relatively high. Therefore, we conducted a study on the transmission mode of rabies and effective suppression measures. First, a mathematical model of the transmission mode of rabies was constructed by drawing relevant data from the officially published data in China and research data from previously published papers. According to this model, the study fit the present number of patients with rabies and predicts the future number of patients with rabies. In addition, we studied the effectiveness of different rabies suppression measures.

In the process of data fitting, through the observation and calculation of each parameter, the BRN of rabies in China was determined to be about 1.069. This figure indicates that the current rabies response is not ideal and cannot effectively prevent the occurrence of rabies. In this paper, when analyzing the effects of each parameter on the BRN, it was found that the higher the number of wild dogs killed, the smaller the number of wild dogs born. Therefore, the value of the BRN can be reduced to less than 1 by killing wild dogs. Controlling the number of births of domestic dogs can also effectively control the value of the BRN. Therefore, the management of domestic dogs in China needs to be strengthened, especially the number of births of domestic dogs, which requires strict control. Although an increase in the degree of immunity of wild dogs can effectively reduce the BRN, the effect of this parameter on the BRN is small and cannot reduce the BRN to less than 1. Improving the degree of immunity of domestic dogs can reduce the value of the BRN to less than 1. The research results in this paper are similar to the results of research conducted by others on how rabies suppression measures affect the number of patients with rabies [14].

Conclusion

Due to the large number of patients with rabies in China and the severity of this condition, this paper investigated the transmission mode of rabies and rabies suppression measures. By counting the number of patients with rabies in China for many years and using software to fit the data, our model forecast an increase in the number of patients with rabies in China in the next few decades. In addition, this paper studied the effect of various measures on reducing the number of patients with rabies. Ultimately, controlling the birth rate of domestic dogs and wild dogs as well as increasing the degree of immunity in domestic dogs are all relatively effective rabies suppression measures. The BRN of rabies in China is still greater than 1. Therefore, in terms of rabies control, China still needs more research, policy formulation, and grassroots implementation.

Although the research in this paper obtained relatively considerable results, there are still some limitations: (1) The number of wild dogs in China is relatively high, and effective sterilization measures have not been implemented for wild dogs. Therefore, this problem was not considered during the construction of the model used in this paper. (2) The research in this paper is still in the theoretical stage, and further studies are required to determine whether it is effective in practice. Therefore, future research will be based on the results of this paper, and will observe whether the results are correct and feasible in practice.

Conflicts of Interest

None declared.

References

1. Beier KT, Kim CK, Hoerbelt P, Hung LW, Heifets BD, DeLoach KE, et al. Rabies screen reveals GPe control of cocaine-triggered plasticity. Nature 2017 Sep 21;549(7672):345-350 [FREE Full text] [doi: 10.1038/nature23888] [Medline: 28902833]
2. Wu S. A Traffic Motion Object Extraction Algorithm. Int J Bifurcation Chaos 2016 Jan 14;25(14):1540039. [doi: 10.1142/s0218127415400398]
3. Cauchemez S, Bourhy H. Improving the provision of rabies post-exposure prophylaxis. The Lancet Infectious Diseases 2019 Jan;19(1):12-13. [doi: 10.1016/s1473-3099(18)30606-6]
4. Adrien J, Georges Y, Augustin PD, Monroe B, Gibson AD, Fenelon N, Haiti-Rabies Field Response Team. Notes from the Field: A Multipartner Response to Prevent a Binational Rabies Outbreak - Anse-à-Pitre, Haiti, 2019. In: MMWR Morb Mortal Wkly Rep. Notes from the Field: A Multipartner Response to Prevent a Binational Rabies Outbreak?Anse-à-Pitre, Haiti, 2019. Morbidity and Mortality Weekly Report; Aug 16, 2019:707-709.
5. Wu S, Wang M, Zou Y. Bidirectional cognitive computing method supported by cloud technology. Cognitive Systems Research 2018 Dec;52:615-621. [doi: 10.1016/j.cogsys.2018.07.035]
6. Birhane MG, Cleaton JM, Monroe BP, Wadhwa A, Orciari LA, Yager P, et al. Rabies surveillance in the United States during 2015. J Am Vet Med Assoc 2017 May 15;250(10):1117-1130. [doi: 10.2460/javma.250.10.1117] [Medline: 28467751]
7. Wu S. Nonlinear information data mining based on time series for fractional differential operators. Chaos 2019 Jan;29(1):013114. [doi: 10.1063/1.5085430] [Medline: 30709142]
8. Kim EJ, Jacobs MW, Ito-Cole T, Callaway EM. Improved Monosynaptic Neural Circuit Tracing Using Engineered Rabies Virus Glycoproteins. Cell Rep 2016 Apr 26;15(4):692-699 [FREE Full text] [doi: 10.1016/j.celrep.2016.03.067] [Medline: 27149846]
9. Lee C, Hwang HS, Lee S, Kim B, Kim JO, Oh KT, et al. Rabies Virus-Inspired Silica-Coated Gold Nanorods as a Photothermal Therapeutic Platform for Treating Brain Tumors. Adv Mater 2017 Apr;29(13):201605563. [doi: 10.1002/adma.201605563] [Medline: 28134459]
10. Wu S, Wang M, Zou Y. Research on internet information mining based on agent algorithm. Future Generation Computer Systems 2018 Sep;86:598-602. [doi: 10.1016/j.future.2018.04.040]
11. National Association of State Public Health Veterinarians, Compendium of Animal Rabies Prevention Control Committee, Brown CM, Slavinski S, Ettestad P, Sidwa TJ, et al. Compendium of Animal Rabies Prevention and Control, 2016. J Am Vet Med Assoc 2016 Mar 01;248(5):505-517. [doi: 10.2460/javma.248.5.505] [Medline: 26885593]
12. Pieracci EG, Brown JA, Bergman DL, Gilbert A, Wallace RM, Blanton JD, et al. Evaluation of species identification and rabies virus characterization among bat rabies cases in the United States. J Am Vet Med Assoc 2020 Jan 01;256(1):77-84. [doi: 10.2460/javma.256.1.77] [Medline: 31841089]
13. Monroe BP, Yager P, Blanton J, Birhane MG, Wadhwa A, Orciari L, et al. Rabies surveillance in the United States during 2014. J Am Vet Med Assoc 2016 Apr 01;248(7):777-788. [doi: 10.2460/javma.248.7.777] [Medline: 27003019]
14. Wu S, Liu J, Liu L. Modeling method of internet public information data mining based on probabilistic topic model. J Supercomput 2019 Jun 19;75(9):5882-5897. [doi: 10.1007/s11227-019-02885-8]

Abbreviations

BRN: basic reproductive number

Edited by Z Du; submitted 09.03.20; peer-reviewed by X Fu, S Wang, F Yao; comments to author 21.03.20; revised version received 26.03.20; accepted 26.03.20; published 27.05.20.

Please cite as:
Huang Y, Li M
Application of a Mathematical Model in Determining the Spread of the Rabies Virus: Simulation Study
JMIR Med Inform 2020;8(5):e18627
URL: http://medinform.jmir.org/2020/5/e18627/
doi:10.2196/18627
PMID:32459185

Original Paper

Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation

Yihao Huang[1,2], MD; Mingtao Li[3], PhD, MD

[1]School of Computer and Information Technology, Shanxi University, Taiyuan, China

[2]Complex Systems Research Center, Shanxi University, Taiyuan, China

[3]College of Mathematics, Shanxi University of Technology, Taiyuan, China

Corresponding Author:
Mingtao Li, PhD, MD
College of Mathematics
Shanxi University of Technology
79, Yingze West St
Taiyuan, 030024
China
Phone: 86 13403459876
Email: mingtaoli@sohu.com

Abstract

Background: Brucella is a gram-negative, nonmotile bacterium without a capsule. The infection scope of Brucella is wide. The major source of infection is mammals such as cattle, sheep, goats, pigs, and dogs. Currently, human beings do not transmit Brucella to each other. When humans eat Brucella-contaminated food or contact animals or animal secretions and excretions infected with Brucella, they may develop brucellosis. Although brucellosis does not originate in humans, its diagnosis and cure are very difficult; thus, it has a huge impact on humans. Even with the rapid development of medical science, brucellosis is still a major problem for Chinese people. Currently, the number of patients with brucellosis in China is 100,000 per year. In addition, due to the ongoing improvement in the living standards of Chinese people, the demand for meat products has gradually increased, and increased meat transactions have greatly promoted the spread of brucellosis. Therefore, many researchers are concerned with investigating the transmission of Brucella as well as the diagnosis and treatment of brucellosis. Mathematical models have become an important tool for the study of infectious diseases. Mathematical models can reflect the spread of infectious diseases and be used to study the effect of different inhibition methods on infectious diseases. The effect of control measures to obtain effective suppression can provide theoretical support for the suppression of infectious diseases. Therefore, it is the objective of this study to build a suitable mathematical model for brucellosis infection.

Objective: We aimed to study the optimized precontrol methods of brucellosis using a dynamic threshold–based microcomputer model and to provide critical theoretical support for the prevention and control of brucellosis.

Methods: By studying the transmission characteristics of Brucella and building a Brucella transmission model, the precontrol methods were designed and presented to the key populations (Brucella-susceptible populations). We investigated the utilization of protective tools by the key populations before and after precontrol methods.

Results: An improvement in the amount of glove-wearing was evident and significant (P<.001), increasing from 51.01% before the precontrol methods to 66.22% after the precontrol methods, an increase of 15.21%. However, the amount of hat-wearing did not improve significantly (P=.95). Hat-wearing among the key populations increased from 57.3% before the precontrol methods to 58.6% after the precontrol methods, an increase of 1.3%.

Conclusions: By demonstrating the optimized precontrol methods for a brucellosis model built on a dynamic threshold–based microcomputer model, this study provides theoretical support for the suppression of Brucella and the improved usage of protective measures by key populations.

(JMIR Med Inform 2020;8(5):e18664) doi:10.2196/18664

KEYWORDS

brucellosis; dynamic model; protective measures; precontrol methods

Introduction

Infectious diseases enter the human body through pathogens such as bacteria, fungi, or viruses, causing bodily damage or even death. In serious cases, infectious diseases cause large-scale transmission of diseases among the population [1]. Furthermore, the number of diseases transmitted from animals to humans is increasing at an alarming rate. Viruses mutate over time, and some mutations make them more suited to living in the current environment and the physiological states of various hosts [2]. More than 60% of the infectious diseases in humans are caused by animals and more than half of animal diseases can be transmitted to humans, which underscores the need for researchers to study such diseases [3]. The infectious disease investigated in this study is brucellosis, which is a zoonotic disease caused by *Brucella* [4].

Brucella is a gram-negative, nonmotile bacterium without a capsule. It is an aerobic intracellular parasite that can reduce nitrates. *Brucella* has a strong ability to adapt to the environment, which makes it able to tolerate dryness and cold temperatures and survive in meat or dairy products for up to 2 months. However, *Brucella* is not heat-resistant and is killed by boiling water. Common disinfectants need several hours to destroy *Brucella*.

In terms of transmission, *Brucella* passes through the digestive tract, the respiratory tract, the skin, and mucous membranes. By eating *Brucella*-contaminated food or contacting animals or animal secretions and excretions infected with *Brucella*, humans can become infected with *Brucella* [5]. Once infected with *Brucella*, the infected person will go through the acute and chronic stages of brucellosis. At first, the infected person is in the acute stage, at which time they will have symptoms of other systemic diseases. During this stage, brucellosis is relatively easily cured; however, the infection is often not effectively diagnosed as brucellosis at this point. After the infected person enters the chronic stage, brucellosis is very difficult to cure [6].

At present, most of the studies on the transmission of brucellosis are in single populations. When studying the process of infection, transmission between humans is ignored, and only direct infection is considered. However, the transmission process of brucellosis is complicated, and its transmission form is not the same in different regions and between populations. Due to its asymptomatic characteristics, an outbreak of brucellosis is often synchronized between humans and animals. Investigating the spread of brucellosis would help researchers understand and prevent the onset and large-scale transmission of brucellosis. Researchers have built a model for the transmission of brucellosis between humans and flocks. The effects of different control methods on the transmission of brucellosis are known, which aids researchers in designing more effective methods to inhibit the spread of brucellosis [7]. Other researchers have built a mixed model of brucellosis based on sheep-bovine-human transmission. This transmission model revealed that disinfection and immunization are the two most effective inhibitory measures against the transmission of brucellosis [8]. This study explores the transmission of brucellosis among goats to provide theoretical support for the suppression of brucellosis.

By understanding the effect that control measures have on transmission suppression, we can create theoretical support for the suppression of infectious diseases [9]. The objective of this study was to construct a mathematical model of brucellosis infection. By studying the transmission characteristics of *Brucella* and building a *Brucella* transmission model, the precontrol methods for key populations (*Brucella*-susceptible populations) were designed. We then determined the utilization rate of protective tools by key groups before and after a presentation on precontrol methods. This study provides theoretical support for the suppression of *Brucella* and protective measures for key populations.

Methods

Proposed Algorithm

Model Construction

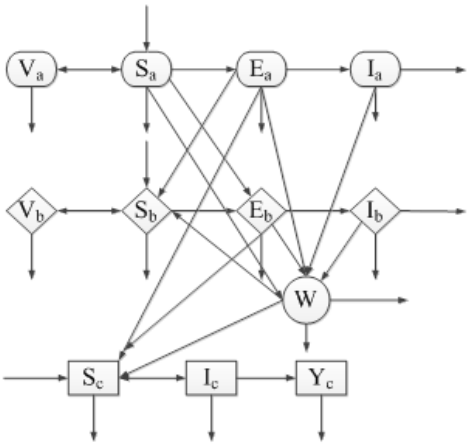
Based on the characteristics of *Brucella* transmission from goat flock to humans, we built a model of *Brucella* transmission that divides the flock into two groups: ordinary ewes and other goats. In addition, we divided the ordinary female flock further into susceptible goats, infected goats, and vaccinated goats. The human population was divided into two groups: susceptible and infected people. Additionally, infected people were divided into acute and chronic infections. Models were built based on the characteristics of *Brucella* transmission. The following hypotheses were made: (1) There is no route of infection among humans. *Brucella* infections in humans come from direct or indirect contact with goats. (2) Goats and humans have natural deaths. (3) The number of human births and the number of goat flocks are considered, as are the natural death of bacteria and bacterial death caused by disinfection. The transmission of brucellosis is shown in Figure 1. The diamonds represent the ordinary ewes, the rounded rectangles represent other goats, and the rectangles represent the human population.

In Figure 1, S_a, V_a, E_a, and I_a represent ordinary ewes that are susceptible, vaccinated, inapparently infected, and isolated positive infections, respectively,. S_b, V_b, E_b, and I_b represent susceptible, vaccinated, inapparently infected, and isolated positive infections of other goats, respectively. W represents *Brucella* in the environment, while S_c, I_c, and Y_c represent susceptible, acute, and chronic human populations, respectively. As shown in Figure 1, *Brucella* infection among susceptible human populations can be caused by *Brucella* in the environment and inapparently infected goats in the flock. *Brucella* in the environment is caused by the inapparently infected goats and the isolated positively infected goats. Additionally, *Brucella* dies naturally in the environment. Therefore, the differential equation of *Brucella* transmission can be obtained, as shown in Equations 1-12.

In these equations, A is the input constant value, b is the probability of each infection, m is the conversion rate of the young to mature goats, d is the production rate, t is the ratio of the adult and young infection rates in the flock, which should

be between 0 and 1 combining with the actual situation, a is the mortality rate of infected sheep due to brucellosis, k is the release rate of infected bacteria per unit time, and d is the Brucella mortality rate in the entire goat flock.

Figure 1. Diagram of the transmission of Brucellosis. S_a: susceptible ewe; V_a: immunized ewe; E_a: recessive infected ewe; I_a: isolated infected ewe; S_b: sheep with hepatitis B; V_b: immunized sheep; E_b: recessive infected sheep; I_b: isolated infected sheep; W: environmental Brucella; S_c: susceptible human population; I_c: acutely infected human population; Y_c: chronically infected human population.



Model Dynamics

Since the last 3 equations are independent of the previous 9, only the first 9 equations are included when considering model dynamics. Within equations 1-9, an equilibrium point free of brucellosis can be found, represented by the equation below:

.

The resulting positive invariant set of the research system is as follows:

Numerical Simulation

Based on the number of brucellosis cases reported on the internet, a hypothetical estimation and numerical simulation process were performed. The values used in the numerical simulation are described here. First, 2-3 years is the inventory time of ordinary ewes, and the number of ewes is about 4.2 million; therefore, the average removal rate of ordinary ewes d_b is 0.4, and the supplement amount of ordinary ewes A_b is 1.68 million. According to the actual data, the production rate of the flock is about 60%. Therefore, the removal rate of other goats is 0.6, and the supplement amount is 1.976 million. Second, according to population data, the natural death rate in the human population is 5.68%; thus, the supplement rate of the human population is estimated, and the supplement amount is about 9,150. Third, according to the existing data, the culling rate of infected goats can reach 0.15; therefore, the positive detection rate of ordinary ewes and other goats was set to 0.15 for this model. Additionally, according to the average 1-month survival time of affected goats, the culling rate of infected ewes and other goats is set to 12.

Precontrol Methods of Brucellosis

Studies have shown that the daily behavior of susceptible populations (that is, those with high frequency of contact with animals) is the key to whether brucellosis can be effectively transmitted, and that the irregular daily behaviors of susceptible populations greatly increase the rate of *Brucella* infection; thus, it is important for them to protect themselves during daily exposure [10]. Due to the generally low level of education of the key populations in this study, they lack self-protection awareness and knowledge of protective measures to varying degrees. In their daily contact with animals or related products, they cannot achieve comprehensive and universal protection from *Brucella* infections.

The precontrol methods in this study included giving lectures on prevention, control, and health education to the experimental population, so that they would have a comprehensive understanding of *Brucella*. This laid the foundation for additional precontrol works. Subsequently, *Brucella* bacterial preventive tools (such as gloves, masks, and disinfectants) were distributed to the key populations and the relevant preventive training was delivered, followed by a question and answer session that answered the key populations' questions about *Brucella* protection. This strengthened their protection awareness and led them to change their daily habits and use of protective tools. The experimental population involved in this study was the staff on a farm. All personnel were included in the experiment and signed informed consent forms.

Follow-up visits or phone calls were made to infected patients to understand whether they are used to the corresponding prevention and control behaviors, and whether their prevention and control behaviors are correct. Additionally, any questions that susceptible populations and infected patients have about brucellosis should be answered promptly to reduce panic and strengthen the effect of precontrol methods.

Statistical Methods

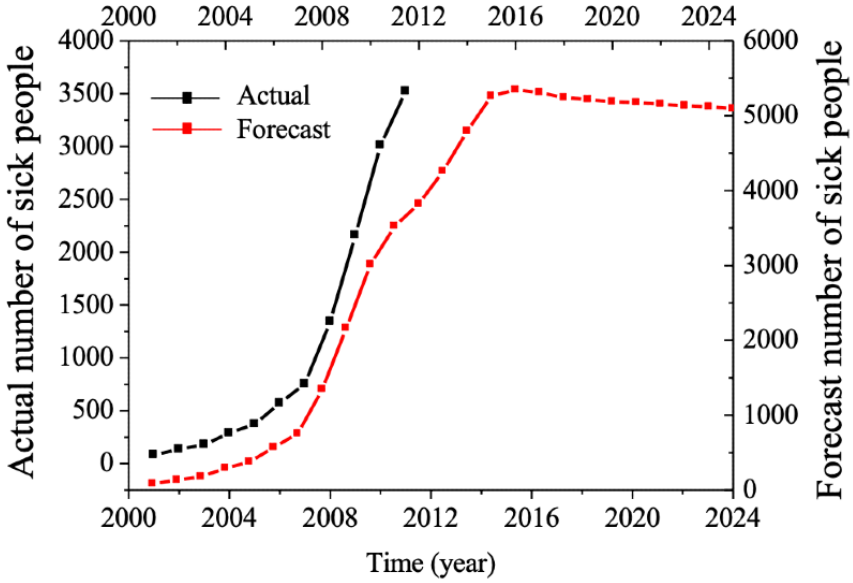
SPSS 22.0 software (IBM) was used to statistically analyze the data obtained in this study and make corresponding statistical descriptions. The Chi-square test was used to statistically analyze the clinical characteristics of the diseased populations with different disease forms and different *Brucella* contact history. $P<.05$ indicates that the difference is statistically significant. For comparison between multiple groups, the Chi-square test was used, and the check level was a=2a/k(k-1), where k is the number of grouping groups, and P<a represents that the differences are statistically significant.

Results

Results of the Numerical Simulation

The numerical simulation process was performed according to the actual data, which enabled us to generate the data shown in Figure 2. As shown in Figure 2, the numerical simulation suggested that the number of people with brucellosis will be stable in the future.

Figure 2. Fitting and short-term prediction of Brucellosis cases. This figure shows the relationship between the number of brucellosis cases and time in the region, and the appropriate number of measures required to obtain a forecast of the number of future brucellosis cases.



Precontrol Results of Brucellosis

The mathematical model of this study suggests that humans are infected with *Brucella* through direct or indirect contact with animals. Therefore, to avoid *Brucella* infection, it is necessary to first isolate humans from *Brucella*, a process in which utilizing many protective tools is the most critical step.

The distribution of brucellosis in the study population is shown in Figure 3, demonstrating that the majority of brucellosis patients are men, and the difference in gender distribution is statistically significant ($P<.001$). In addition, most brucellosis patients are aged 30-59 years. One possible reason for this is that the experimental population was the staff of a farm, where most staff are in that age range.

Figure 4 shows the proportion of different symptoms experienced in different stages of the disease. As shown in Figure 4, the clinical manifestations of patients with brucellosis in the acute stage are mostly fever, pain in muscles and joints, sweating, and fatigue. The above clinical phenomena are significantly more common in patients in the acute stage than those in the chronic stage. The difference is statistically significant ($P=.006$). Conversely, liver and spleen enlargement occurred at both disease stages, and the difference was not statistically significant. However, the probability of testicular enlargement in patients was significantly higher in the acute stage than in the chronic stage; the difference was statistically significant ($P=.003$). Finally, the likelihood of lymphadenopathy in patients in the chronic stage is significantly higher than that in patients in other stages ($P=.005$).

Figure 5 shows the utilization of protective tools by the key populations before and after the intervention. As shown in the figure, the utilization of protective tools by the key populations improved significantly following the intervention. Many people in the key populations now wear gloves, masks, rubber shoes, hats, and work clothes; wash their hands frequently; and disinfect animal shelters. The improvement in glove-wearing was the most striking, as it increased from 51.01% before the precontrol methods to 66.22% after the precontrol methods, an increase of 15.21%. The difference was statistically significant ($P<.001$). However, hat-wearing did not improve significantly ($P=.08$). It increased from 57.3% before the precontrol methods to 58.6% after the precontrol methods, an increase of just 1.3%.

Figure 3. The number of people with brucellosis in different stages by gender (A) and age (B).

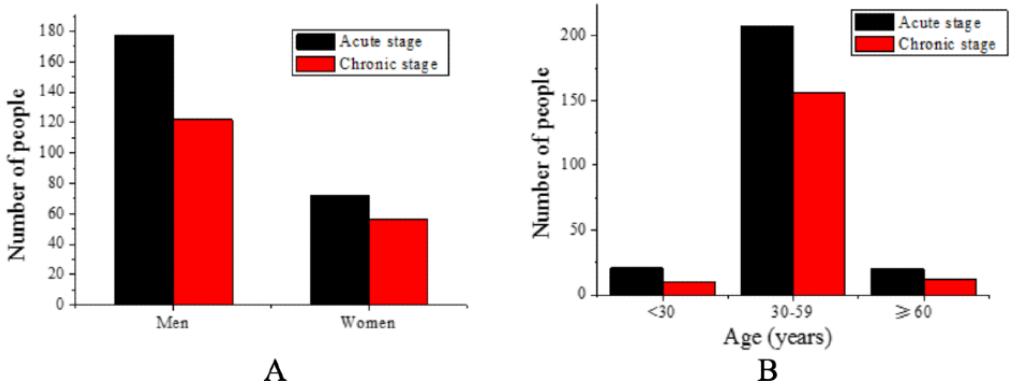


Figure 4. Clinical symptoms of patients at different stages.

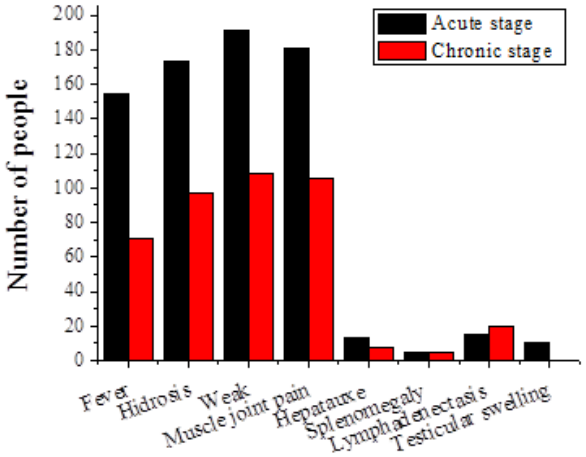
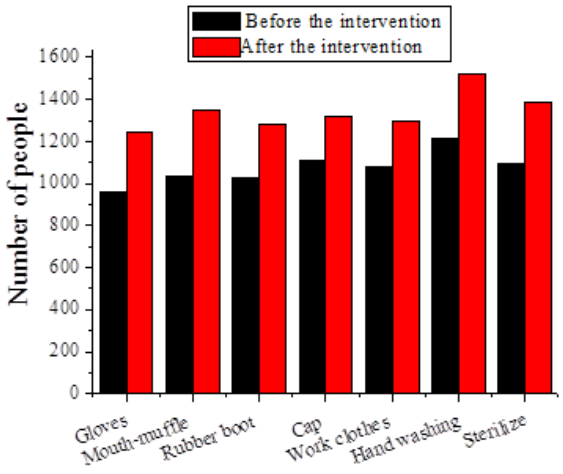


Figure 5. Utilization of protective tools by the key populations.



Discussion

Overview

Brucella has a strong ability to adapt to the environment. It can tolerate dryness and cold temperatures, and can survive in meat and dairy products for up to 2 months. Presently, it is not possible to completely remove the infection sources of brucellosis. Additionally, it is not currently possible to develop a practical and effective vaccine for brucellosis. Studies have shown that the daily behavior of susceptible populations (that is, those with a high frequency of contact with animals) is the key to whether brucellosis is transmitted further. Indeed, the irregular daily behaviors of susceptible populations could greatly increase the rate of *Brucella* infection; thus, it is important that they protect themselves when in contact with animals, consistent

with previously published research [11] on the suppression of brucellosis transmission.

In this study, we first designed a transmission model of *Brucella* in animals and humans. It found that human infection by *Brucella* is through direct or indirect contact with animals. Therefore, to avoid *Brucella* infection, it is necessary to be isolated from animals infected with *Brucella*. The utilization of various protective equipment and tools is critical. Prior to the precontrol methods and information dissemination, due to the generally low level of education in the key populations studied, people lacked knowledge about protective measures to varying degrees. When in daily contact with animals or related products, people in these key populations could not achieve comprehensive and universal protection from *Brucella* infections.

After the precontrol methods in this study, the study populations improved their use of protective tools compared to before the precontrol methods. The utilization of several protective tools improved to varying degrees. In particular, the improvement in glove-wearing was the most striking, as it increased from 51.01% before the precontrol methods to 66.22% after the precontrol methods, an increase of 15.21%. The difference was statistically significant (P<.001). However, hat-wearing did not improve significantly (P=.95), as it increased from 57.3% before the precontrol methods to 58.6% after the precontrol methods, an increase of 1.3%. A possible reason is that people in the key populations do not notice the protective effects of wearing a hat during daily contact with animals; thus, they do not pay attention to wearing hats as protection. The research results also reflected the age distribution and level of education of the key populations; their acceptance of information is slow, and their acceptance of new knowledge is low. However, in general, the utilization of protective tools by the key populations improved significantly. Most people in the key populations wear gloves, masks, rubber shoes, hats, and work clothes; wash their hands frequently; and disinfect animal shelters.

Conclusion

This study researched brucellosis transmission and the effects of precontrol methods on protective equipment usage by key populations. The results of our numerical simulation indicated that the incidence of brucellosis is projected to become stable, without a major increase or decrease. After the precontrol methods, the utilization of protective tools by the key populations improved significantly. Most people in the key populations wear gloves, masks, rubber shoes, hats, and work clothes; wash their hands frequently; and disinfect animal shelters. This study achieved our objectives, but there are still some deficiencies in the research process. Due to the limitation of time, this study failed to analyze the prevalence of brucellosis in key populations after the precontrol methods. In a future study, we aim to research precontrol methods and the subsequent number of brucellosis infections.

Conflicts of Interest

None declared.

References

1. Tuon FF, Gondolfo RB, Cerchiari N. Human-to-human transmission of Brucella - a systematic review. Trop Med Int Health 2017 May 09;22(5):539-546 [FREE Full text] [doi: 10.1111/tmi.12856] [Medline: 28196298]
2. Mustafa AS, Habibi N, Osman A, Shaheed F, Khan MW. Species identification and molecular typing of human Brucella isolates from Kuwait. PLoS ONE 2017 Aug 11;12(8):e0182111. [doi: 10.1371/journal.pone.0182111]
3. Shaofei W, Mingqing W, Yuntao Z. Research on internet information mining based on agent algorithm. Future Generation Computer Systems 2018 Sep;86:598-602. [doi: 10.1016/j.future.2018.04.040]
4. Carbonero A, Guzmán L, García-Bocanegra I, Borge C, Adaszek L, Arenas A, et al. Seroprevalence and risk factors associated with Brucella seropositivity in dairy and mixed cattle herds from Ecuador. Trop Anim Health Prod 2017 Sep 26;50(1):197-203. [doi: 10.1007/s11250-017-1421-6]
5. Wu S, Wang M, Zou Y. Bidirectional cognitive computing method supported by cloud technology. Cognitive Systems Research 2018 Dec;52:615-621. [doi: 10.1016/j.cogsys.2018.07.035]
6. Costa Franco MM, Marim F, Guimarães ES, Assis NRG, Cerqueira DM, Alves-Silva J, et al. Triggers a cGAS-Independent STING Pathway To Induce Host Protection That Involves Guanylate-Binding Proteins and Inflammasome Activation. JI 2017 Dec 04;200(2):607-622. [doi: 10.4049/jimmunol.1700725]
7. Wu S, Zhang Q, Chen W, Liu J, Liiu L. Research on trend prediction of internet user intention understanding and public intelligence mining based on fractional differential method. Chaos, Solitons & Fractals 2019 Nov;128:331-338. [doi: 10.1016/j.chaos.2019.07.034]
8. Köhler, S. Ouahrani-Bettache, J. Y. Winum. Brucella suis carbonic anhydrases and their inhibitors: Towards alternative antibiotics. Journal of enzyme inhibition and medicinal chemistry; 2017:683-687.
9. El-Diasty M, Wareth G, Melzer F, Mustafa S, Sprague L, Neubauer H. Isolation of Brucella abortus and Brucella melitensis from Seronegative Cows is a Serious Impediment in Brucellosis Control. Veterinary Sciences 2018 Mar 09;5(1):28. [doi: 10.3390/vetsci5010028]
10. Wu S. Nonlinear information data mining based on time series for fractional differential operators. Chaos 2019 Jan;29(1):013114. [doi: 10.1063/1.5085430]

11. Rhyan, M. Garner, T. Spraker. Brucella pinnipedialis in lungworms Parafilaroides sp. and Pacific harbor seals Phoca vitulina richardsi: proposed pathogenesis. Diseases of aquatic organisms; 2018:87-94.

Edited by K Kalemaki; submitted 11.03.20; peer-reviewed by Y Lin, Z Wang, R Zhang; comments to author 19.03.20; revised version received 21.03.20; accepted 23.03.20; published 27.05.20.

Please cite as:
Huang Y, Li M
Optimization of Precontrol Methods and Analysis of a Dynamic Model for Brucellosis: Model Development and Validation
JMIR Med Inform 2020;8(5):e18664
URL: https://medinform.jmir.org/2020/5/e18664
doi:10.2196/18664
PMID:32459180

Original Paper

Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development

Yue Liu[1], MA

The First People's Hospital of Fuyang, Hangzhou, Zhejiang Province China, Hangzhou, China

Corresponding Author:
Yue Liu, MA
The First People's Hospital of Fuyang
Zhejiang Province China
429 Beihuan Road
Fuyang District
Hangzhou, 311400
China
Phone: 86 13588381028
Email: hzliuyue1982@163.com

Abstract

Background: The incidence of diabetes is increasing in China, and its impact on national health cannot be ignored. Smart medicine is a medical model that uses technology to assist the diagnosis and treatment of disease.

Objective: The aim of this paper was to apply artificial intelligence (AI) in the diagnosis of diabetes.

Methods: We established an AI diagnostic model in the MATLAB software platform based on a backpropagation neural network by collecting data for the cases of integration and extraction and selecting an input feature vector. Based on this diagnostic model, using an intelligent combination of the LabVIEW development platform and the MATLAB software-designed diabetes diagnosis system with user data, we called the neural network diagnostic module to correctly diagnose diabetes.

Results: Compared to conventional diagnostic procedures, the system can effectively improve diagnostic efficiency and save time for physicians.

Conclusions: The development of AI applications has utility to aid diabetes diagnosis.

(JMIR Med Inform 2020;8(5):e18682) doi:10.2196/18682

KEYWORDS

artificial intelligence; diabetes; neural network

Introduction

Artificial Intelligence and Smart Medicine

Artificial intelligence (AI) has meaningful benefits for human beings; its promotion in the development of medical technology can enable machines, algorithms, and big data to serve human health needs. In China, the incidence of diabetes is on the rise, and its impact on national health cannot be ignored. Diabetes is a disease in which the body metabolizes proteins, fats, sugars, and other substances in the blood, liver, and other organs because it does not properly produce or use insulin.

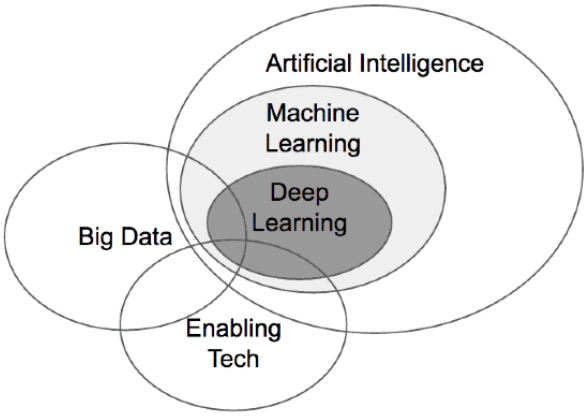
Smart medicine is a medical model that uses AI technology to assist diagnosis and treatment. In the future, smart medicine will be a core technology to fight disease and prolong human life. The advent of AI has increased the intelligence of computers, enabling them not only to “learn” expert medical knowledge but also to simulate the thinking and reasoning of physicians regarding patients. Therefore, computers can provide reliable diagnoses and treatment plans. AI can process huge amounts of data rapidly. Through the study of big data, AI can discover rules and summarizes their differences with regularity to diagnose diseases. Based on this background, in this paper, we introduced AI technology to an auxiliary diabetes diagnosis system. We used an AI neural network algorithm to establish an auxiliary diagnosis model and judge the type of diabetes based on physiological parameters input by a user. This may aid the diagnosis of diabetes in under-resourced areas or by inexperienced physicians.

Theoretical Basis

AI is an area of research and development involving the simulation, extension, and expansion of cutting-edge science and interdisciplinary intelligence theory, methods, techniques,

and applications. The famous British scientist Alan Turing is known as the father of AI; in 1950, he published a famous paper [1], "Computing Machinery and Intelligence," which provided the first definition of AI from a behaviorist point of view and proposed the concept of a "thinking machine" as well as the "Turing test" to determine whether a machine is intelligent. Many human activities, such as problem solving, riddle solving, online shopping, preparation of plans, writing computer programs, and even driving a car, require intelligence. If a machine can perform such tasks, the machine can be considered to possess artificial intelligence. Machine learning is a type of AI that uses algorithms to analyze data, learn from the analysis, and draw inferences or make predictions. As an important mathematical algorithm model in machine learning, neural networks have been successfully applied to many practical problems that are difficult for computers to solve in the fields of automatic control, pattern recognition, medicine, and economics. Neural networks have good intelligence characteristics and strong nonlinear processing ability. The application of neural networks in diabetes analysis and research has great significance. The relationship between neural networks, machine learning, and AI is shown in a Venn diagram in Figure 1 [2].

Figure 1. Venn diagram showing the relationship between neural networks, machine learning, and artificial intelligence.



Overview of Artificial Neural Networks

Figure 2 shows a multilayer forward neural network. The computing nodes in this network are hidden layer and output layer neurons.

Hebb's learning rule adjusts the connection weight (w_{ij}) between neurons according to the principle that when neurons i and j are simultaneously excited, the connection between them must be strengthened:

$$\Delta w_{ij} = \eta u_i(t) u_j(t) \quad (1)$$

Neurocytology has confirmed that this rule is consistent with conditioned reflex theory [a]. $\eta(0<\eta<1)$ represents the proportionality constant of the learning rate and is also called the learning factor or learning step.

The δ learning rule adjusts the connection weight (w_{ij}) between neurons. When the output value of a neuron does not match the expected value, the weight of the neuron must be adjusted according to the difference between the expected value and the actual value [3]:

$$\Delta w_{ij} = \eta[d_i - y_i(t)]y_i(t) \quad (2)$$

This is the minimum mean square error learning rule. It is a special case of the δ learning rule. Its principle is adjustment of the mean square error between the actual output of the neuron and the expected output to the minimum, that is:

(3)

The principle is that "the winner is fully profitable," that is, the neuron in a layer of neurons that produces the largest output value for the input is the "winner." Then, the weights connected to the "winner" can simply be adjusted to bring it closer to the valuation of the input sample pattern:

$$\Delta w_{ij} = \eta[g(x_i) - w_{ij}(t)] \quad (4)$$

A neural network simulates the structure of the human brain; the simplest model of a neural network is a perceptron, which consists of a set of interconnected nodes constituting a chain. Figure 3 shows the structure of a perceptron. It contains 2 types of nodes: several input nodes, consisting of attributes used to represent the input, and an output node that provides an output model. The nodes in a neural network structure are called neurons or units. In a perceptron, the input nodes and output node are connected by a weighted chain.

After the weighted summation of the input by the perceptron, the offset factor t is subtracted, and then the output value $\hat{y}$ is obtained according to the sign of the result. For example, in a perceptron with 3 input nodes, the weight of each node to the output node is 0.3, the paranoid factor t is 0.4, and the output calculation formula of the model is as follows:

(5)

The difference between the input node and the output node of the perceptron is that the input node directly transmits the received value to the output chain without any conversion, while the output node performs a weighted summation on the input and then subtracts the bias term; the symbol then produces an

output according to the result. Equation 6 is a mathematical representation of the output of the perceptron model:

$$\hat{y} = sign(w1x1 + w2x2 + ... + wnxn - t) \quad (6)$$

where w_1, w_2, ... , w_n is the weight of the input chain, x_1, x_2, ... , x_n is the input attribute value, and *sign* is a symbolic function; when the parameter is positive, the activation function of the neuron is output +1, and when the parameter is negative, the activation function of the neuron is output –1 [4]. The perceptron model can also be expressed by Equation 7:

$$\hat{y} = sign(wx - t) \quad (7)$$

where w is the weight vector and x is the input vector.

Figure 2. Schematic of a multilayer forward neural network.

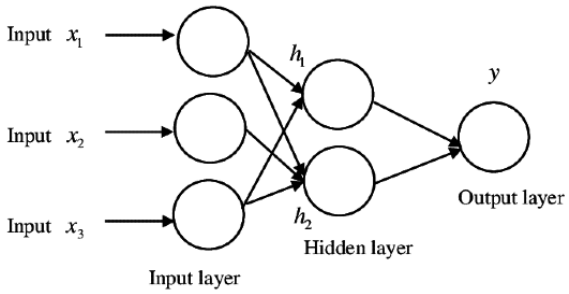
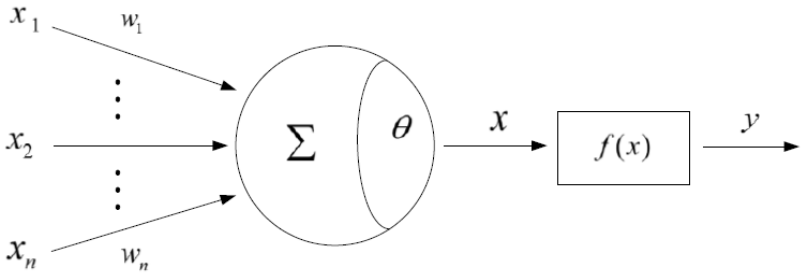


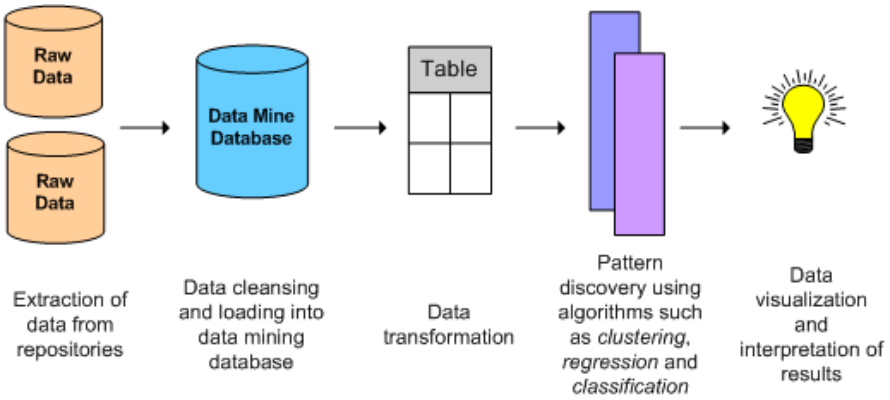
Figure 3. Schematic of a perceptron.



Principles of Data Mining

Data mining is also referred to as data collection or data gathering. The process of data mining involves finding and extracting information and knowledge with potential use value from a large amount of incomplete, noisy, and random practical application data. Data mining is an iterative process. In this process, information discovery is performed manually or intelligently. When exploring and researching an unknown database, it is impossible to predict the knowledge and information it contains. In contrast to traditional data analysis technology, data mining technology can describe the hidden features of many data categories, predict their development trends, and identify useful information with high decision-making value and guiding significance for work and life. As shown in Figure 4, data mining is universally applicable to various types of data [5].

Figure 4. Schematic of the data mining process.

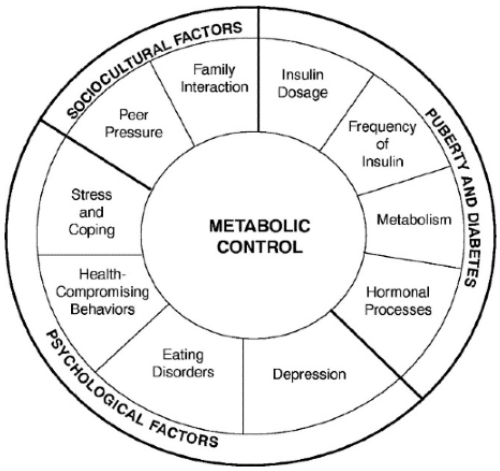


Diabetes Data Processing

Influencing Factors of Diabetes

At present, because the onset of diabetes is extremely complicated, its etiology and pathogenesis have not been fully elucidated. The recognized factors influencing the incidence of diabetes according to many medical report studies and data analyses of diabetic patients are shown in Figure 5.

Figure 5. Influencing factors of diabetes.



Genetic Factors

The onset of diabetes has familial characteristics; if family members in the previous generation have diabetes, the probability of their offspring having diabetes is very high. The genetic factors of diabetes are clear. The prevalence of diabetes in people with a blood relationship to patients with diabetes is 5 times higher than that of people who do not have such a blood relationship [6]. Moreover, the risk of genetic factors in type 1 diabetes is 40% higher than in type 2 diabetes [7]. Medical research has identified a variety of genetic mutations that affect the onset of diabetes. For example, type 1 diabetes is closely related to the DQ site polymorphism in the human leukocyte antigen gene, and multiple DNA sites will affect the onset of the disease. Type 2 diabetes is also affected by mutations of genes such as the insulin-hormone gene and the glucokinase gene [8].

Dietary Factors

When the human body consumes more energy than it expends for a long time, obesity can result. Obesity is the most important risk factor for diabetes. Most patients with type 2 diabetes have other conditions, such as obesity, hypertension, and hyperlipidemia.

Environmental Factors

Environmental factors affecting diabetes mainly include population aging, reduced exercise volume, viral infection, and psychological stress. In populations susceptible to conditions caused by genetic factors, the onset of diabetes may also be affected by environmental factors. Lifestyle changes also greatly contribute to the increasing number of people with type 2 diabetes. In an American Indian tribe in Arizona, the incidence of obese and diabetic patients rose from 0% to 50% due to changes in the original lifestyle of the tribe members [9].

Psychological Factors

Psychological factors are often ignored by researchers; however, in recent years, many clinical medical studies have found that psychological factors have effects on the pathogenesis of diabetes [8]. Most scholars hold the view that when a person remains in a state of mental stress and depression for a long time, excessive stress and negative emotions will cause the body to respond to stress, quickly produce a large number of hormones that can increase blood sugar, and inhibit the production of insulin [3]. Studies have found that when a person's mood fluctuates, the blood glucose levels in their body will rise to varying degrees. If the person remains in this state for a long time, they have a high chance of developing diabetes [5].

Other Factors

There are various additional predisposing factors for diabetes. In addition to the 4 factors mentioned above, comprehensive influences include various physical indicators, such as age, gender, height and weight, blood pressure and blood lipids; adverse living habits, such as smoking and alcohol abuse; and taking estrogen [7].

Diagnostic Criteria of Diabetes

The blood glucose cutoff points for diabetic diagnosis are fasting blood glucose ≥7.0 millimoles per liter (126 milligrams per deciliter); 2 hour oral glucose tolerance test blood glucose ≥11.1 mmol/L (200 mg/dL); and random blood glucose ≥11.1 mmol/L (200 mg/dL) [10].

Patients with typical symptoms of diabetes (eg, frequent drinking, polyuria, polyphagia, weight loss) and who meet any of the above cutoff points can be diagnosed with diabetes. For those with no obvious symptoms, only the first or second cutoff point can be used as a diagnostic condition, and the patient's blood glucose levels must be checked on another day. In impaired glucose regulation, the fasting blood glucose or glucose tolerance test 2 hours after taking glucose exceeds normal values but does not meet the diagnostic criteria for diabetes.

Methods

Collection and Preprocessing of Diabetes Data

We combined the theoretical knowledge of medical data mining to integrate and preprocess medical data from a Top 3 hospital in Lanzhou. The initial fasting blood glucose and 2 hour postprandial blood glucose levels of all patients with diabetes exceeded the standard values. More than 95% of the initial symptoms of patients with type 2 diabetes included thirst, dry mouth, excessive drinking, polyphagia, polyuria, and weight

loss. The inheritance of diabetes was mainly manifested in type 1 diabetes. Type 2 diabetes was mostly caused by acquired habits or endocrine disorders. By comparing health data, it was found that people who smoke and drink perennially are at 50% higher risk for diabetes. Therefore, the above symptoms were used as important indicators for diagnosing patients with diabetes. In the data, 1 indicates yes and 0 indicates no.

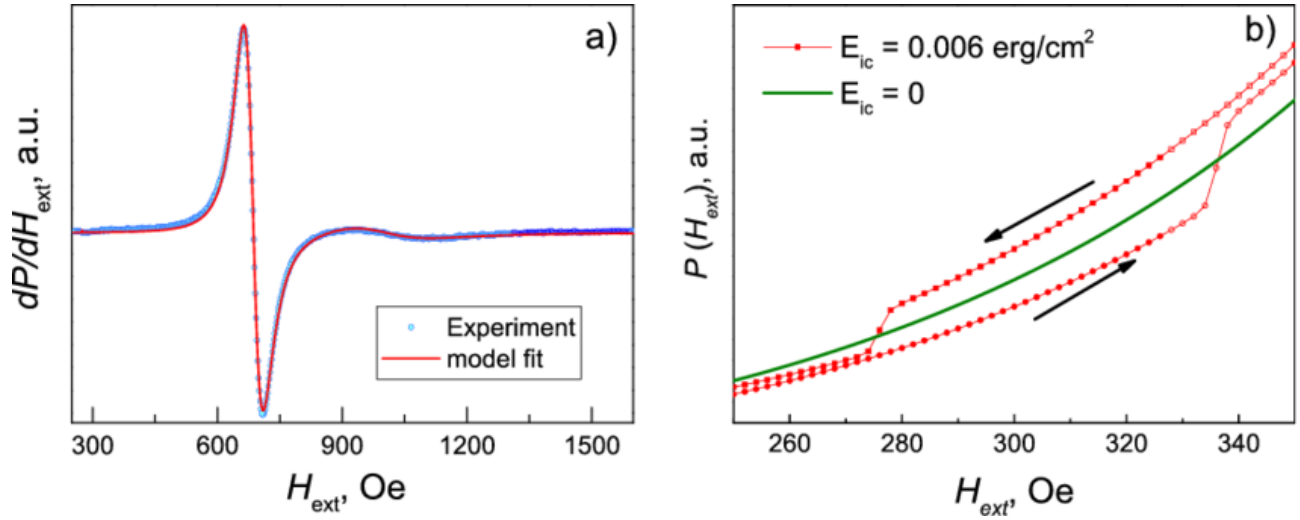
Establishment of the Diagnosis Model and Analysis of the Results

Model Establishment Based on the Backpropagation Neural Network

There is currently no quantified standard for selecting the number of hidden layer units. is usually used to determine the number of hidden layer units. In this study, *n* was taken as 19, *m* as taken as 1, and the range of the numbers of hidden layer units in this model was initially determined.

As can be seen in Figure 6, when the number of hidden layer units is 12, the network and the number of iterations of the training effects are the best. Therefore, in this paper, we selected the number of hidden layer units as 12 to model the backpropagation neural network.

Figure 6. Establishment of the model based on the BP neural network.



Results

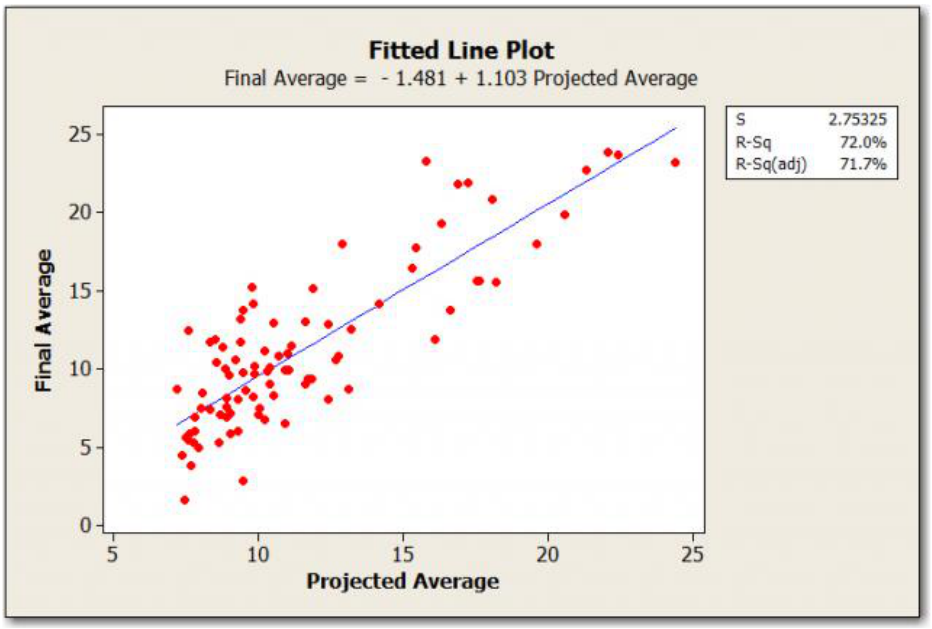
The sample regression coefficients are shown in Figure 7.

The 650 screened groups of patients with diabetes included 395 groups (60.8%) of patients with type 2 diabetes and 255 groups (39.2%) of patients with type 1 diabetes. The training set contained 590 groups in total, including 343 groups (58.1%) of patients with type 2 diabetes and 247 groups (41.9%) of patients with type 1 diabetes. The test set used 60 groups of data, including 39 groups (65%) of patients with type 2 diabetes and 21 groups (35%) of patients with type 1 diabetes.

The outputs of the test set were obtained after training the backpropagation neural network. The results showed that of the 41 type 2 diabetes groups, 38 (93%) were accurately diagnosed, while 3 (7%) were misdiagnosed. For the 19 type 1 diabetes groups, 17 (89%) were accurately diagnosed, while 2 (11%) were misdiagnosed.

Among the test data of the 60 groups, 5 groups of data were misdiagnosed, and the average correct diagnosis rate was 55/60 (92%). This percentage is feasible for diabetic diagnosis.

Figure 7. Sample regression with 12 hidden layer units.



Discussion

Implementation of a Diabetes Diagnosis System

The MATLAB platform (MathWorks, Inc) has a long history of use in research and application of neural networks; a variety of its implementation technologies and methods are now quite mature, especially since the emergence of the neural network toolbox. However, MATLAB has many limitations in practical applications, such as a poor user interface. Therefore, in this paper, we combined LabVIEW software (National Instruments Corporation) with MATLAB; we used the LabVIEW platform to design a graphical user interface while calling MATLAB script nodes and using their functional characteristics to diagnose diabetes.

System Design User Interface

The user login interface was mainly constructed using the while loop, event structure, Boolean function, string control, and conditional structure in LabVIEW. At the same time, the username subvirtual instrument (SubVI) and password decision SubVI were set in the main program of the login system.

When the Boolean space of the front panel account login value changes, the block diagram will check whether the account and password match the data stored in the user VI. A conditional structure is used. When the account and password are correct, the judgment condition is true; thus, the attribute node opens and the user can enter the main program, as shown in Figure 8. When the account and password data are not found in the user VI, the judgment condition is false, and an account or password error occurs. The program was written using both local variables and attribute nodes; thus, it is simple and functional. The password decision SubVI is shown in Figure 9.

In order to adapt to the situation where the user forgets their account name and password when using the diabetes diagnosis system, a password modification VI was designed in the main program. When the user forgets the account name and password, the password can be retrieved through the password recovery interface; Figure 10 shows a block diagram of the password modification function, which includes the conditional structure, Boolean function, and string control.

In the main program, the diabetes diagnosis system is constructed with a tiled sequential structure. There is no sequential structure in text programming; a text programming language can execute statements in order, but the order of sequential execution is changed by loops and conditional structures. In contrast, LabVIEW uses a data flow method involving a multi-threaded parallel structure that drives the programming direction through the data flow. At the same time, we can implement multi-threading without any additional programming, which is an additional advantage of multi-threaded operation. In this operation mode, whether the node can run normally in LabVIEW depends entirely on whether the data are flowing into all the required inputs of the node. LabVIEW data is free to “swim.” It does not necessarily follow the default direction of data flow from left to right. Therefore, the sequence in which the block diagrams run is somewhat random, and the sequence of the LabVIEW block diagrams cannot be decided with complete accuracy.

Figure 8. The main program of the login system.

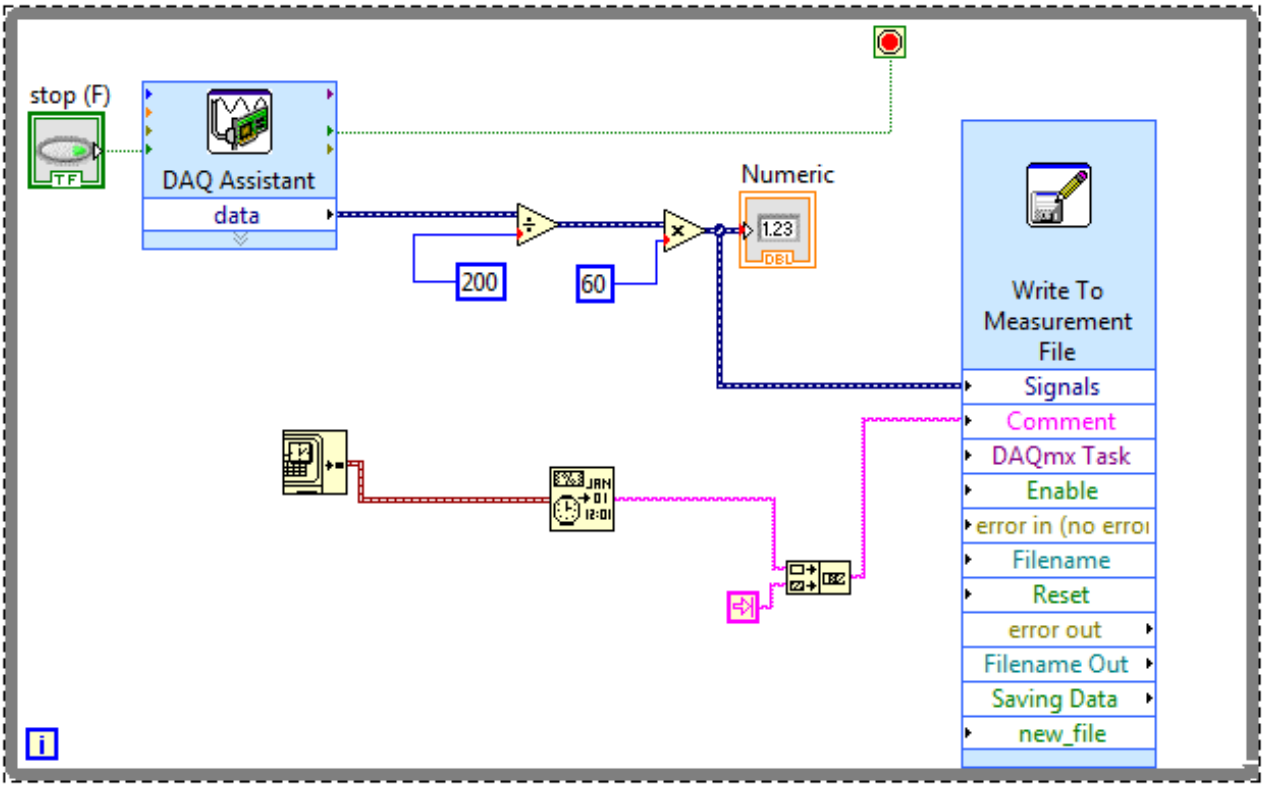


Figure 9. The password decision SubVI.

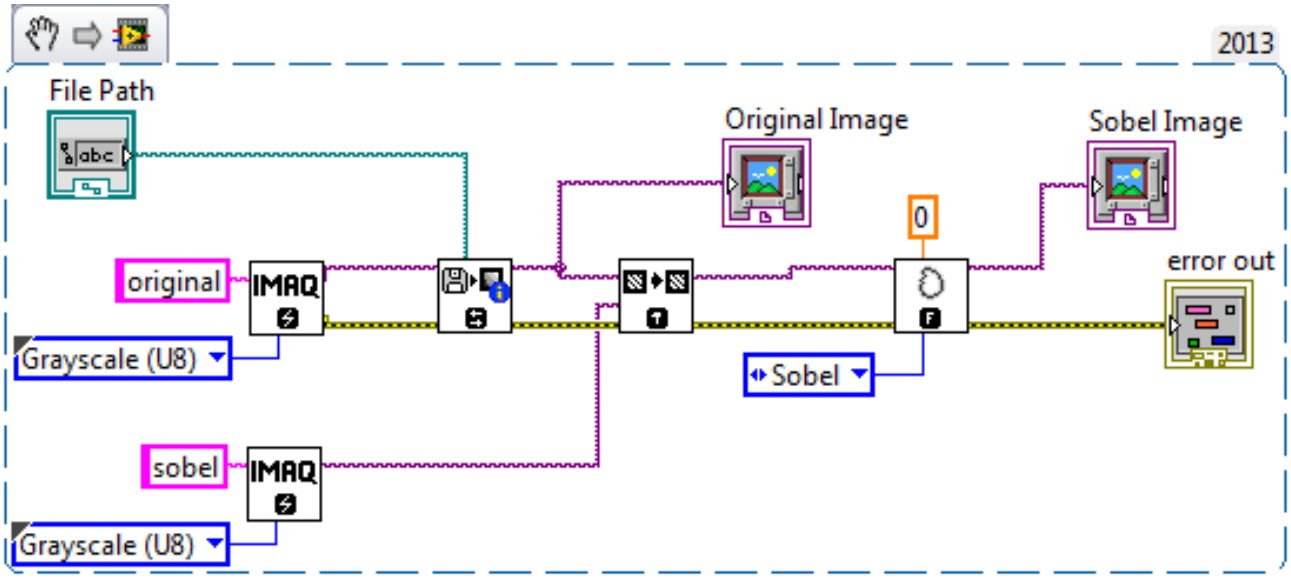
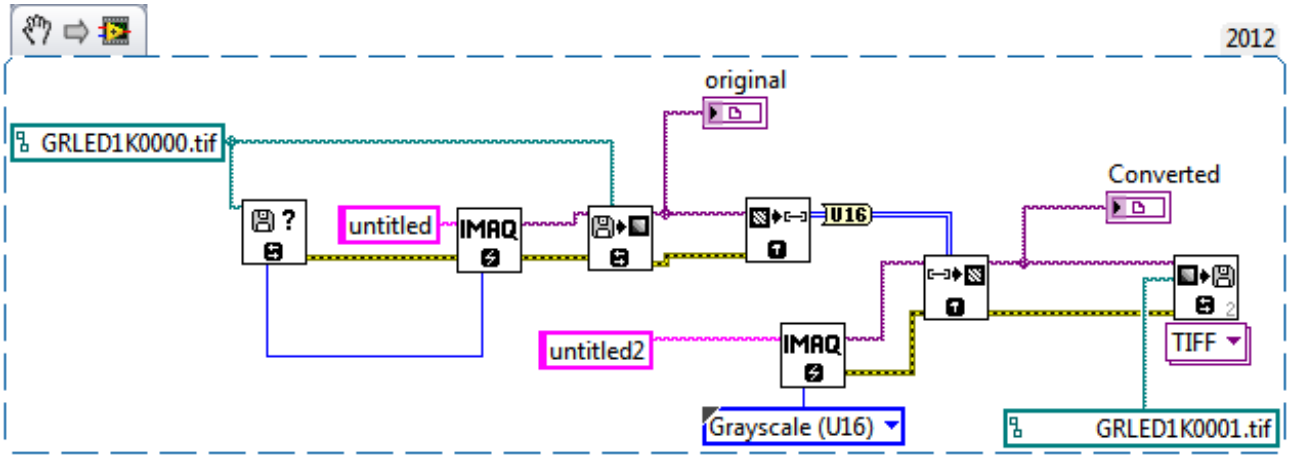


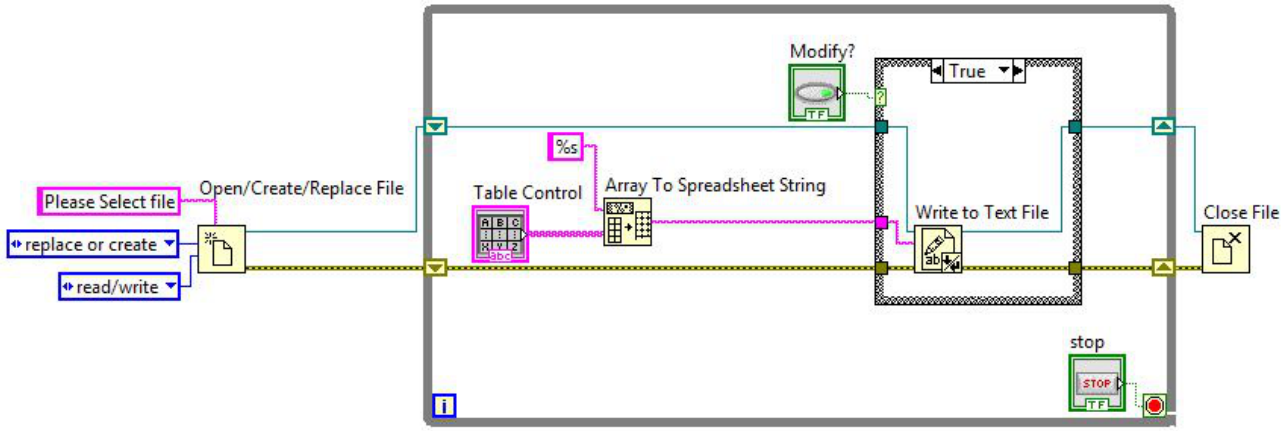
Figure 10. The password modification VI.



Call of the MATLAB Neural Network Diagnostic Module

LabVIEW can relate to other languages and software through interfaces; it can also be extended to multiple target platforms and operating systems to achieve user requirements for different numerical calculations and analyses. By running programs with MATLAB scripts in LabVIEW and communicating with MATLAB script nodes in MATLAB through LabVIEW, MATLAB's powerful numerical calculation functions can be used in LabVIEW. Based on the above background, in this paper, we used LabVIEW to build a diabetes diagnosis system while modeling neural networks with MATLAB. The locations of the MATLAB script nodes in LabVIEW are shown in Figure 11.

Figure 11. Locations of the MATLAB script nodes in LabVIEW.



Conclusions

In recent years, the development of smart medicine has been increasing in China as well as in other countries. Breakthroughs in key technologies such as image recognition, depth of learning, and neural networks have resulted in a new round of developments in AI. On the other hand, with the continuous improvement of human living standards, the incidence of diabetes in our country has also continued to rise; in reality, medical resources are unevenly distributed, with insufficient numbers of physicians, less experienced community physicians, and other issues. More efficient allocation of medical resources reduces misdiagnosis due to lack of experience; this paper proposes combining AI technology with diabetes diagnosis to build a diagnostic system that helps physicians diagnose diabetes.

Conflicts of Interest

None declared.

References

1. Turing AM. Computing Machinery and Intelligence. Mind 1950 Oct;59(236):433-460. [doi: 10.1093/mind/LIX.236.433]

2. Lawrence DR, Palacios-González C, Harris J. Artificial Intelligence. Camb Q Healthc Ethics 2016 Mar 09;25(2):250-261. [doi: 10.1017/s0963180115000559]
3. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol 2018 Jun;71(23):2668-2679. [doi: 10.1016/j.jacc.2018.03.521]
4. Shaofei W, Mingqing W, Yuntao Z. Research on internet information mining based on agent algorithm. Future Gener Comp Sy 2018;86:598-602. [doi: 10.1016/j.future.2018.04.040]
5. Lopes BT, Eliasy A, Ambrosio R. Artificial Intelligence in Corneal Diagnosis: Where Are We? Curr Ophthalmol Rep 2019 Jul 9;7(3):204-211. [doi: 10.1007/s40135-019-00218-9]
6. Kontoangelos K, Papageorgiou CC, Raptis AE, Tsiotra P, Boutati E, Papadimitriou GN, et al. The role of oxytocin, cortizol, homocysteine and cytokines in diabetes mellitus and their association with psychological factors. Arch Hellen Med 2014;31(1):7-22 [FREE Full text]
7. Dong G, Qu L, Gong X, Pang B, Yan W, Wei J. Effect of Social Factors and the Natural Environment on the Etiology and Pathogenesis of Diabetes Mellitus. Int J Endocrinol 2019;2019:8749291 [FREE Full text] [doi: 10.1155/2019/8749291] [Medline: 31341475]
8. Martinez-Millana A, Bayo-Monton JL, Argente-Pla M, Fernandez-Llatas C, Merino-Torres JF, Traver-Salcedo V. Integration of Distributed Services and Hybrid Models Based on Process Choreography to Predict and Detect Type 2 Diabetes. Sensors (Basel) 2017 Dec 29;18(1) [FREE Full text] [doi: 10.3390/s18010079] [Medline: 29286314]
9. Ravussin E, Valencia ME, Esparza J, Bennett PH, Schulz LO. Effects of a traditional lifestyle on obesity in Pima Indians. Diabetes Care 1994 Sep;17(9):1067-1074. [doi: 10.2337/diacare.17.9.1067] [Medline: 7988310]
10. Mathur P, Burns ML. Artificial Intelligence in Critical Care. Int Anesthesiol Clin 2019;57(2):89-102. [doi: 10.1097/AIA.0000000000000221] [Medline: 30864993]

Abbreviations

AI: artificial intelligence
SubVI: subvirtual instrument
VI: virtual instrument

Edited by K Kalemaki; submitted 12.03.20; peer-reviewed by И Иванова, D Yang, L Huang; comments to author 19.03.20; revised version received 21.03.20; accepted 22.03.20; published 27.05.20.

Please cite as:
Liu Y
Artificial Intelligence–Based Neural Network for the Diagnosis of Diabetes: Model Development
JMIR Med Inform 2020;8(5):e18682
URL: http://medinform.jmir.org/2020/5/e18682/
doi:10.2196/18682
PMID:32459183

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org